Published in final edited form as:
Anal Chem. 2014 February 18; 86(4): 2221–2228. doi:10.1021/ac403879d.

Fragment Formula Calculator (FFC): Determination of Chemical Formulas for Fragment Ions in Mass Spectrometric Data

André Wegner[*,†,‡], **Daniel Weindl**[†], **Christian Jäger**[†], **Sean C. Sapcariu**[†], **Xiangyi Dong**[†], **Gregory Stephanopoulos**[‡], and **Karsten Hiller**[†]

[†]Luxembourg Centre for Systems Biomedicine, University of Luxembourg, 7, avenue des Hauts-Fourneaux, L-4362 Esch-Belval, Luxembourg

[‡]Department of Chemical Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, 56-439, Cambridge, MA 02140, United States

Abstract

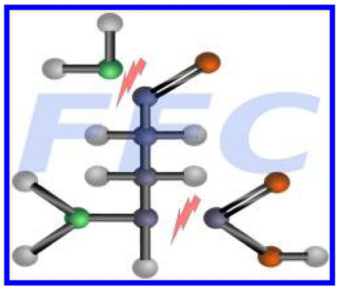
The accurate determination of mass isotopomer distributions (MID) is of great significance for stable isotope-labeling experiments. Most commonly, MIDs are derived from gas chromatography/electron ionization mass spectrometry (GC/EI-MS) measurements. The analysis of fragment ions formed during EI, which contain only specific parts of the original molecule can provide valuable information on the positional distribution of the label. The chemical formula of a fragment ion is usually applied to derive the correction matrix for accurate MID calculation. Hence, the correct assignment of chemical formulas to fragment ions is of crucial importance for correct MIDs. Moreover, the positional distribution of stable isotopes within a fragment ion is of high interest for stable isotope-assisted metabolomics techniques. For example, ^{13}C-metabolic flux analyses (^{13}C-MFA) are dependent on the exact knowledge of the number and position of retained carbon atoms of the unfragmented molecule. Fragment ions containing different carbon atoms are of special interest, since they can carry different flux information. However, the process of mass spectral fragmentation is complex, and identifying the substructures and chemical formulas for these fragment ions is nontrivial. For that reason, we developed an algorithm, based on a systematic bond cleavage, to determine chemical formulas and retained atoms for EI derived fragment ions. Here, we present the fragment formula calculator (FFC) algorithm that can calculate chemical formulas for fragment ions where the chemical bonding (e.g., Lewis structures) of the intact molecule is known. The proposed algorithm is able to cope with general molecular rearrangement reactions occurring during EI in GC/MS measurements. The FFC algorithm is able to integrate stable isotope labeling experiments into the analysis and can automatically exclude candidate formulas that do not fit the observed labeling patterns.[1] We applied the FFC algorithm to create a fragment ion repository that contains the chemical formulas and retained carbon atoms of a wide range of trimethylsilyl and tertbutyldimethylsilyl derivatized compounds. In total, we

[*]**Corresponding Author**andre.wegner@uni.lu..

The authors declare no competing financial interest.

ASSOCIATED CONTENT
Supporting Information
Additional information as noted in text. This material is available free of charge via the Internet at http://pubs.acs.org.

report the chemical formulas and backbone carbon compositions for 160 fragment ions of 43 alkylsilyl-derivatives of primary metabolites. Finally, we implemented the FFC algorithm in an easy-to-use graphical user interface and made it publicly available at http://www.ffc.lu.



Stable isotope labeling experiments (SLE) have emerged as an important tool in metabolic engineering and systems biology.[2] Of key concern for SLE is the accurate assessment of isotopomer distributions of cellular metabolites by gas chromatography/mass spectrometry (GC/MS) and nuclear magnetic resonance (NMR).[3] While NMR lacks sensitivity, it provides detailed positional information. In contrast, GC/MS allows a sensitive determination of isotopic enrichment but only provides limited positional information. Over the last years, powerful techniques such as metabolic flux analysis (MFA) have been developed to determine metabolic fluxes in biological systems based on the mass isotopomer distributions (MID) of small molecules.[4–6] MFA has been applied to many biomedical and biotechnological problems.[7–11] Usually, MIDs for mass spectral fragment ions can be calculated only if the chemical formula of the specific fragment ion is known, except if a special experimental setup is used.[12] Hence, most often only the information of the molecular ion peaks are used for MID measurements. However, electron ionization (EI)-based mass spectrometry leads to complex mass spectra, caused by the fragmentation of the analyzed compound. The analysis of fragment ions, which contain only specific parts of the original molecule, can provide valuable information on the positional isotopic enrichment within the molecule of interest. This positional distribution of the label is of high interest for ^{13}C-MFA. In addition, based on the applied derivatization method, the molecular ion might not be visible at all and fragment ions have to be analyzed instead. An important consideration is that the process of assigning a chemical structure to a fragment ion from a known molecular ion structure is time-consuming, even for an expert.[13]

In this work, we propose a novel method for the determination of chemical formulas and retained atoms for EI fragment ions based on the two-dimensional (2D) structure of a compound in combination with the measured mass spectrum. In general, there are two ways to deal with EI-based fragmentation: a rule-based in silico prediction or a combinatorial approach. Rule-based algorithms, such as ACD/MS Fragmenter or Mass Frontier,[14] rely on fragmentation mechanisms derived from molecules where the fragmentation is known, assuming that similar structures will fragment the same way. However, small changes in structure can lead to a significantly different fragmentation mechanism.[13] Furthermore, the rule-based approach fails for molecules where no similar fragmentation mechanism is known. A combinatorial approach usually is based on a systematic bond cleavage. For that, a cleavage cost is assigned to each bond to find the substructure with minmal costs. Finding the correct cost function, however, is challenging. For example, MetFrag[15] uses bond-dissociation energies, whereas FiD[16] uses standard bond energies. One drawback of current rule-based and combinatorial approaches is that they can only capture simple hydrogen rearrangements but fail for more complex rearrangements.

Here, we present a universal method to determine chemical formulas for fragment ions without a priori knowledge about the fragmentation mechanisms, taking advantage of the combinatorial aspect of the problem. A method based on a similar idea has been proposed for high-resolution tandem mass spectrometry.[16] However, our method is designed for MS data with nominal masses, as produced by most GC/MS instruments with a quadrupole mass analyzer, which are routinely used in many laboratories. In contrast to high-resolution MS data determining chemical formulas for nominal masses is algorithmically more challenging, because there are many possible permutations of elemental compositions that cannot easily be ruled out. In addition, our algorithm is able to cope with molecular rearrangements, which occur frequently in EI measurements.

THEORETICAL BACKGROUND

The fragmentation of gas phase ions is a complex and often hard-to-predict process. A detailed description can be found elsewhere.[13] Although the whole fragmentation process can be very complex, there are only a few basic types of reactions that break or form chemical bonds: (1) σ-ionization, immediately breaks a bond (affecting mostly hydrocarbons); (2) α-cleavage, a new bond is formed from a radical site and an adjacent bond is homolytically cleaved; (3) charge-induced heterolytic cleavage, cleavage of a bond next to a charge-site; (4) rearrangements, migrations of atoms or groups of atoms (see Figure 1); (5) displacement of atoms or groups of atoms; and (6) eliminations.

Graph theory has been extensively used in the fields of biology and chemistry. To model the fragmentation of a molecule, we will apply its graph-theoretical representation to determine chemical formulas of mass spectrometric fragment ions. On the basis of the fragmentation rules described above, a fragment ion is always composed of a subset of atoms of the original molecule. By using graph theory, the problem of assigning a chemical formula to a fragment ion can, therefore, be broken down to finding a subgraph H of G, assuming the graph G represents the structure of the molecular ion.

A graph is an ordered pair $G = (V,E)$ where V is a set of vertices (or nodes) and E a set of edges. Each element of E contains a pair (u,v), elements of V. The term labeled graph refers to a graph G, where a label is assigned to the set of vertices and edges. Formally, this is expressed by the two functions f_V: $V \rightarrow A$ for the set of vertices and f_E: $V \times V \rightarrow B$ for the set of edges. If B is an ordered set (e.g., real numbers) then the graph is called weighted and the value $f_E\ (u,v)$ is called the weight of the edge from u to v. A connected component C of a graph G has every pair of vertices joined by a path. A connected graph consists of one connected component. The removal of a set of edges, which disconnects the graph, is called a cut. A subgraph of $G = (V,E)$ is a graph $H = (W,F)$, where W is a subset of V, and F is a subset of E, and all edges in F have their end points in W.

ALGORITHM

We model a molecule as an undirected, connected, and labeled graph $G = (V, E, f_{VA}, f_{VB}, f_{VC}, f_{ED})$, where V is the set of vertices corresponding to the atoms and E is the set of undirected edges corresponding to the bonds between the atoms. The function $f_{VA}: V \rightarrow A$ assigns each atom an element (e.g., carbon, hydrogen, etc.), $f_{VB}: V \rightarrow B$ assigns each atom an index, and $f_{VC}: V \rightarrow C$ assigns each atom the atomic mass according to the chemical element. The function $f_{ED}: V \times V \rightarrow D$ assigns each bond an order (single, double, or triple). The mass of the molecular ion corresponds to the sum of the masses of all vertices:

$$W(G) = \sum_{v \in V} f_{VC}(v) \quad (1)$$

The underlying idea of this algorithm is that the fragmentation process usually only breaks a few bonds within the molecule. This can be simulated by removing a defined number of edges within the molecular graph. In terms of graph theory this means to induce a cut of a certain size in in the graph. This can leave the graph G disconnected. The resulting connected components $C = \{C_1, ..., C_n\}$ of the subgraph H each have a molecular mass:

$$W(C_i) = \sum_{v \in V(C_i)} f_{VC}(v) \quad (2)$$

Since the mass (m) of the fragment ion is determined by mass spectrometry, the chemical formula of this fragment ion corresponds to a combination of connected components of H, in which molecular masses $W(C_i)$ sum up to m. Figure 2 illustrates this process. The resulting subgraph (representing the chemical composition), which can be composed of several connected components, does not necessarily represent the chemical structure because the formation of new bonds (e.g., fragmentation rule 4) is not modeled. However, the number and position of atoms of the intact compound retained in this fragment ion is uncovered.

So far, we have relied on the assumption that the correct edges are deleted from the graph. There are two unknowns, the number and the position of edges to be deleted. To define the minimal number of edges to delete from the graph (cut size), necessary to model the fragmentation, it is mandatory to take the fragmentation rules (as stated in Theoretical Background) into consideration. Fragmentation types 1–3 cleave one bond without forming new σ-bonds, 4 and 5 cleave one bond while forming a new one, 6 cleaves two bonds while forming a new one. Therefore, to describe an α-cleavage or a σ-ionization, clearly a cut size of one is sufficient. To simulate a simple elimination or a rearrangement, which is equivalent to deleting one edge in the graph, a cut size of one is also necessary. For the combination of a more complex rearrangement and an α-cleavage (as depicted in Figure 1), a cut size of three is necessary. To capture both the single and the combined fragmentations, the algorithm is designed to work with a defined maximum cut size. The cut size starts at one and subsequently increases until it reaches the defined maximum cut size.

One way to find the correct edges to delete from the graph is to select those edges that are most likely to break. For example, low-energy bonds can be assumed to break more easily.

Although this is correct, additional rules are needed to describe rearrangements. Another more straightforward way is to delete all possible combinations of edges of a certain cut size. Certainly this includes the correct edges but at the same time increases the number of possible results enormously. If the number of edges is given by *n* and the cut size by *k*, then the number of *k* distinct elements of n is given by the binomial coefficient:

$$\binom{n}{k}=\frac{n!}{k!\cdot(n-k)!} \quad (3)$$

For example, the graph of the molecule *N*,*O*-bis-(trimethylsilyl)-glycine with the molecular formula $C_8H_{21}NO_2Si_2$ has 33 edges. The number of possible distinct edge sets to delete for a cut size of 3 is then 5456.

To find the correct edges, the resulting fragment formulas for each of these possibilities have to be ranked according to a score. At best, this score is linked to the measured mass spectrum. One elegant way to do so is to predict the theoretical mass spectrum of the determined fragment formula and calculate a spectrum similarity score to the measured mass spectrum of this fragment ion. A mass spectrum can be theoretically predicted by using the natural stable isotopic distribution of elements and statistical theory.[18] For elements that only have one naturally occurring stable isotope of significant abundance, the distribution of isotopes can be predicted by a binomial distribution:

$$m_i=\frac{n!}{i!\cdot(n-i)!}\cdot p_0^{n-1}\cdot p_1^{i} \quad (4)$$

where n is the total number of atoms, i the number of atoms containing the heavier isotope (e.g., ^{13}C), p_0 the natural abundance of the lighter isotope [e.g., $p(^{12}\mathrm{C}) = 0.989$] and p_1 the natural abundance of the heavier isotope [e.g., $p(^{13}\mathrm{C}) = 0.01$]. In case an element has several natural occurring isotopes, the distribution of those isotopes within a molecule can be predicted by a multinomial distribution:

$$m_i=\frac{n!}{a_1!\cdot a_2!\cdot\ldots\cdot a_k!}\cdot p_0^{a_0}\cdot p_1^{a_1}\cdot\ldots\cdot p_k^{a_k} \quad (5)$$

where *n* is the total number of atoms, a_0 to a_k the number of atoms containing the respective isotope, and p_0 to p_k the natural abundances of those isotopes.

Reducing Algorithmic Complexity

For GC/MS, compounds are usually derivatized prior to analysis. For example, active protons in functional groups (hydroxyl-, carboxyl-, thiol-, amino groups, etc.) can be replaced with a trimethylsilyl (TMS) or *tert*-butyldimethylsilyl (TBDMS) group. This makes compounds more volatile and less reactive but at the same time increases the computational complexity of finding the correct chemical formula of a fragment ion. In the case of stable isotope labeling experiments, the interest lies normally only in labeling patterns for atoms of the original (underivatized) molecule. As a consequence, the information obtained from the loss of atoms originating from the derivatization reagent used is often redundant. For example, when TMS derivatization is used, a $[M - 15]^+$ fragment is often present in the mass spectrum, originating from the loss of a methyl group from the derivatized part of the molecule. Depending on the number of TMS groups within the molecule, there are several possibilities for the position of the lost methyl group. With regard to the calculation of chemical formulas, however, the position of this methyl group is not relevant and computational time can thus be saved. For that reason, we divide the molecular graph into atoms belonging to the original molecule (backbone atoms) and atoms originating from the derivatization reagent used. Subsequently, nonbackbone edges (edges that are not connected to at least one backbone atom) are grouped based on the atoms that would be lost if this edge is deleted (Figure 3). For example, all edges are grouped together where their removal would lead to the loss of one hydrogen. This reduces the number of distinct edges significantly, thereby decreasing the combinatorial complexity for the problem of finding the correct chemical formula. Additionally, this allows the user to follow the fate of specific atoms in the molecular ion by selecting them as backbone atoms.

Another advantage which makes the proposed algorithm capable of modeling rearrangements is the use of connected components. Fragment ions resulting from a rearrangement reaction are often composed of two or more disjoint substructures of the molecular ion. Identifying these substructures is computationally challenging, as their number grows enormously with the number of atoms. However, in our algorithm, the number of these substructures is limited by the number of connected components within the molecular graph, making the proposed algorithm also applicable for larger molecules.

Constraining/Weighting the Result Set

One problem of finding a chemical formula through a combinatorial- instead of a rule-based approach is the high number of possible results. One way to remove redundant results is to consider only results where either the molecular formula or the composition of backbone atoms changes. In other words, results with the same chemical formula but different nonbackbone atoms are ignored (as stated above). Although this shrinks the result set considerably, it still leaves a fair amount of candidate formulas. For that reason, the FFC program allows for the addition of a spectrum of a stable isotope labeling experiment to the analysis. Labeled fragments are automatically detected, and MIDs for those fragments are calculated in order to determine the number of labeled atoms within this fragment. Candidate formulas that do not fit the labeling pattern are directly excluded from the result set.

MATERIAL AND METHODS

Details can be found in the Supporting Information.

IMPLEMENTATION

FFC has been developed in C++ and Qt4 and is based on the publicly available MetaboliteDetector,[19] NTFD,[20] and the ICBM algorithm.[21] All graph-based calculations are done using the LEMON graph library,[22] available at http://lemon.cs.elte.hu.

RESULTS AND DISCUSSION

We first validated the predictive capabilities of FFC by identifying the chemical formulas for 35 fragment ions of 13 *tert*-butyldimethylsilyl derivatized amino acids. These manually curated formulas have been published previously by Antoniewicz.[1] The mass spectra as well as the 2D structures were obtained from the NIST 08 library. An overview of all fragment ions tested is depicted in Table 1 of the Supporting Information. We tested whether FFC can not only predict the correct formula but also the correct position of retained backbone carbon atoms, which is very important for MFA. We considered a predicted formula as correct when the candidate with the lowest number of broken bonds matched the formula proposed by Antoniewicz. If there were multiple formulas resulting from the same number of broken bonds, we selected the formula with the highest spectrum similarity score. For the composition of backbone carbon atoms, selecting the correct solution is more challenging because candidates with different backbone carbon atoms but the same formula will have the same spectrum similarity score. For this reason, we only considered the prediction of backbone carbon atoms present to be correct if there was a unique solution. Overall, FFC was able to correctly predict 34 out of 35 chemical formulas and 30 out of 35 backbone carbon compositions. In the case of threonine, the formula and the carbon atoms for the fragment ion at *m/z* 376 were predicted incorrectly. However, when we used a spectrum measured using our Agilent 5975C MSD, both the formulas and the carbon atoms were predicted correctly. Apparently, the spectrum similarity score is dependent on the quality of the spectra used and how close it reflects the theoretical distribution of naturally occurring isotopes. The number of correctly predicted formulas is slightly higher compared to the number of backbone carbon compositions because of similar structural groups within the 2D structure of certain molecules. For example, aspartate 3TBDMS and glutamate 3TBDMS both have two carboxyl groups; for the ions at *m/z* 390 (Asp) and 330 (Glu), it is not clear which of these two groups is cleaved off. The chemical formula, however, is the same. In case of leucine and isoleucine, the side chains have the same chemical formula (C_4H_9) as the *tert*-butyl group and, therefore, have the same mass (*m/z* 57) and cannot be distinguished by our algorithm. The top two ranked candidate formulas for ions at *m/z* 200, 274, and 302 of *N*,*O*-bis(dimethyl-*tert*-butylsilyl)-leucine are depicted in Figure 4. For the ion at *m/z* 302, there are two equally ranked candidate formulas, resulting from either an α-cleavage of the *tert*-butyl group or the side chain. Interestingly, Antoniewicz showed with a stable isotope labeling experiment that two fragments with the same chemical formula are overlapping for this ion. He found significant M + 2 and M + 6 mass isotopomer abundances when using U–^{13}C-leucine. This suggests that both backbone carbon atom compositions predicted by FFC are legitimate. For the ion at *m/z* 274, again there are two equally ranked candidate formulas, resulting from either a loss of a *tert*-butyl and the carbonyl group or the loss of the side chain and the carbonyl group. However, when using U–^{13}C-leucine, only the M + 5 peak is abundant, suggesting that five of the six carbon backbone atoms are still present in this fragment.[1] This result can be explained by the rearrangement mechanism depicted in Figure 1. The retro-Diels–Alder-like rearrangement occurs only if the N-terminal *tert*-butyl is lost in a previous fragmentation step, leading to the loss of the carbonyl group. As these two candidate formulas cannot be distinguished solely from unlabeled spectra (unless an expert in the field is looking at it), a stable isotope labeling experiment should be performed to determine which formula is correct. For ion 200, the correct formula is $C_{11}H_{26}NSi$, resulting from an α-cleavage between the carbon of the carboxyl group and the adjacent carbon atom. The second best hit with the formula $C_{11}H_{24}OSi$ has a slightly higher spectrum similarity score of 0.999866 (compared to 0.999819) but needs the higher number of broken bonds, which is very unlikely from a chemical point of view. In our analysis, the correct chemical formula for each fragment ion was always present in the list of results. However, as with most prediction algorithms, a critical look at the result is necessary in order to pull out those that are most chemically relevant.

Next, we applied the FFC program to determine the chemical formulas and carbon backbone compositions of a wide range of trimethylsilyl- (Tables 1, 2, and 3) and *tert*-butyldimethylsilyl- (Table 2 of the Supporting Information) derivatized compounds of central carbon metabolism. In this article, we report a fragment ion repository that includes the chemical formulas and the retained carbon atoms for 160 fragment ions of 43 compounds. The retained carbon backbone compositions of all compounds can be found in the Supporting Information. We manually curated these formulas and verified them with labeled reference spectra. For that, we generated fully ^{13}C-labeled yeast extracts as described in the Materials and Methods section of the Supporting Information. These labeled spectra can be imported in the FFC program, and results that do not fit the labeling pattern are directly removed from the result set. We additionally validated the TMS spectra with deuterated *N*-methyl-*N*-(trimethyl-d_9-silyl)trifluoroacetamide (MSTFA-d_9) as a derivatization reagent. In conclusion, we present a high quality fragment ion repository that can help researchers to analyze stable isotope-labeling experiments. For example, the fragment formulas can be used to calculate MIDs, which in turn can be used in combination with the retained carbon atoms to perform ^{13}C-MFA.

The calculation time for this algorithm is dependent on the size of the molecule and the maximum cut size. For small molecules like *N*,*O*-bis-(trimethylsilyl)-glycine the run time is in the range of miliseconds, whereas for bigger molecules like (1*Z*)-*O*-methyloxime-2,3,4,5,6-pentakis-*O*-(trimethylsilyl)-glucose the run time is in the range of seconds on a standard PC.

CONCLUSION

In this article, we present FFC as an algorithm to not only calculate chemical formulas but also retained atoms of a compound in its mass spectrometric fragment ions. Knowing the correct number and position of specific atoms present in a fragment ion is of great significance for MFA. Although only carbon atoms were tracked in the validation experiment, in theory any element's fate (e.g., nitrogen, sulfur, and hydrogen) can be followed with this algorithm. We provide an easy to use software with a user-friendly graphical interface. Due to the combinatorial nature of our approach, it is not necessary to model the fragmentation based on a rule set, such as the preferred site of ionization or the bonds most likely to break. This also allows the calculation of chemical formulas for compounds where no similar fragmentation mechanism is known. However, identical structural groups present in the compound of interest can complicate interpretation when there is ambiguity in the results (e.g., alkanes, sugars, or fatty acids). To further filter out incorrect formulas, FFC can integrate results of a stable isotope labeling experiment to exclude results that do not fit the labeling pattern. In this article, we showed that this algorithm can be successfully applied to a wide range of biochemical compounds by identifying the chemical formulas and carbon backbone combinations for a wide range of compounds.

FFC is freely available under http://www.ffc.lu. Currently, installable packages for Linux (Debian, Red Hat packages), Mac OS, and Windows are provided.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

The authors gratefully thank Kazunori Sawada and Patrick May for their constructive comments. The authors acknowledge financial support from the Fonds National de la Recherche (FNR). Specifically, K.H. and D.W. are funded by the ATTRACT program Metabolomics Junior Group, A.W. is supported by the AFR Grant 1328318, and S.C.S. is supported by the HICE virtual institute.

REFERENCES

1. Antoniewicz, MR. Ph.D. thesis. Massachusetts Institute of Technology; 2006. Comprehensive Analysis of Metabolic Pathways Through the Combined Use of Multiple Isotopic Tracers..
2. Sauer U. Mol. Syst. Biol. 2006; 2:62. [PubMed: 17102807]
3. Antoniewicz MR, Kelleher JK, Stephanopoulos G. Anal. Chem. 2007; 79:7554–7559. [PubMed: 17822305]
4. Antoniewicz MR, Kelleher JK, Stephanopoulos G. Metab. Eng. 2007; 9:68–86. [PubMed: 17088092]
5. Villas-Boas SG, Moxley JF, Akesson M, Stephanopoulos G, Nielsen J. The Biochemical Journal. 2005; 388:669–677. [PubMed: 15667247]
6. Nöh K, Grönke K, Luo B, Takors R, Oldiges M, Wiechert W. Journal of Biotechnology. 2007; 129:249–267. [PubMed: 17207877]
7. Metallo CM, Gameiro P. a. Bell EL, Mattaini KR, Yang J, Hiller K, Jewell CM, Johnson ZR, Irvine DJ, Guarente L, Kelleher JK, Vander Heiden MG, Iliopoulos O, Stephanopoulos G. Nature. 2012; 481:380–384. [PubMed: 22101433]
8. Wegner A, Cordes T, Michelucci A, Hiller K. Current Biotechnology. 2012; 1:88–97.
9. Niklas J, Priesnitz C, Rose T, Sandig V, Heinzle E. Appl. Microbiol. Biotechnol. 2012; 93:1637–1650. [PubMed: 21842438]
10. McGuirk S, Gravel S-P, Deblois G, Papadopoli DJ, Faubert B, Wegner A, Hiller K, Avizonis D, Akavia UD, Jones RG, Giguére V, St-Pierre J. Cancer and Metabolism. 2013; 1:22. [PubMed: 24304688]
11. Michelucci A, Cordes T, Ghelfi J, Pailot A, Reiling N, Goldmann O, Binz T, Wegner A, Tallam A, Rausell A, Buttini M, Linster CL, Medina E, Balling R, Hiller K. Proc. Natl. Acad. Sci. U.S.A. 2013; 110:7820–7825. [PubMed: 23610393]
12. Hiller K, Metallo CM, Kelleher JK, Stephanopoulos G. Anal. Chem. 2010; 82:6621–6628. [PubMed: 20608743]
13. McLafferty FW, Turecek FJ. Chem. Educ. 1994; 71:A54.
14. HighChem Mass Frontier 7.0. Highchem; Bratislava, Slovakia: 2011.
15. Wolf S, Schmidt S, Müller-Hannemann M, Neumann S. BMC Bioinf. 2010; 11:148.
16. Heinonen M, Rantanen A, Mielikäinen T, Kokkonen J, Kiuru J, Ketola RA, Rousu J. Rapid Commun. Mass Spectrom. 2008; 22:3043–3052. [PubMed: 18763276]

17. Stein SE, Scott DR. J. Am. Soc. Mass Spectrom. 1994; 5:859–866. [PubMed: 24222034]

18. Fernandez CA, Des Rosiers C, Previs SF, David F, Brunengraber HJ. Mass Spectrom. 1996; 31:255–262.

19. Hiller K, Hangebrauk J, Jäager C, Spura J, Schreiber K, Schomburg D. Anal. Chem. 2009; 81:3429–3439. [PubMed: 19358599]

20. Hiller K, Wegner A, Weindl D, Cordes T, Metallo CM, Kelleher JK, Stephanopoulos G. Bioinformatics (Oxford, England). 2013; 29:1226–1228.

21. Wegner A, Sapcariu SC, Weindl D, Hiller K. Anal. Chem. 2013; 85:4030–4037. [PubMed: 23514283]

22. Dezso B, Jüttner A, Kovács P. Electronic Notes in Theoretical Computer Science. 2011; 264:23–45.

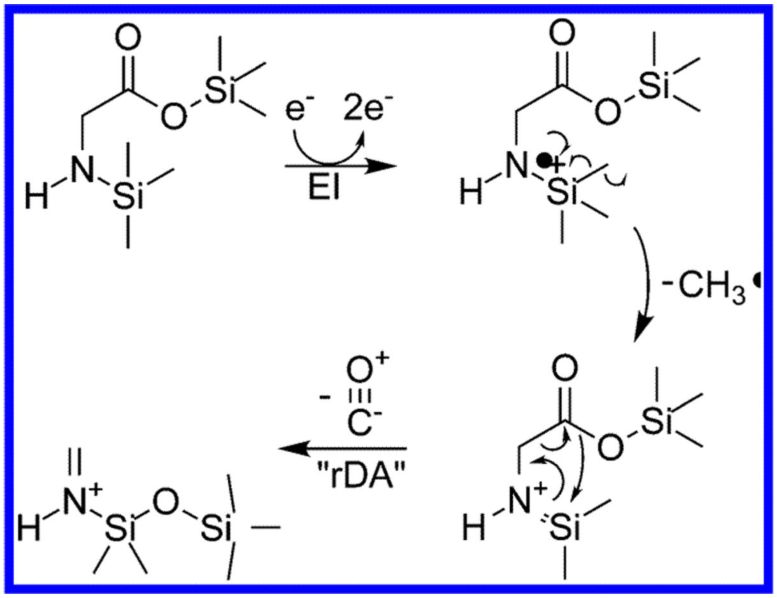


Figure 1.

Proposed fragmentation mechanism of *N*,*O*-bis-(trimethylsilyl)-glycine. After expulsion of a methyl radical by alpha cleavage next to the nitrogen, carbon monoxide loss occurs by a retro-Diels–Alder-like reaction.

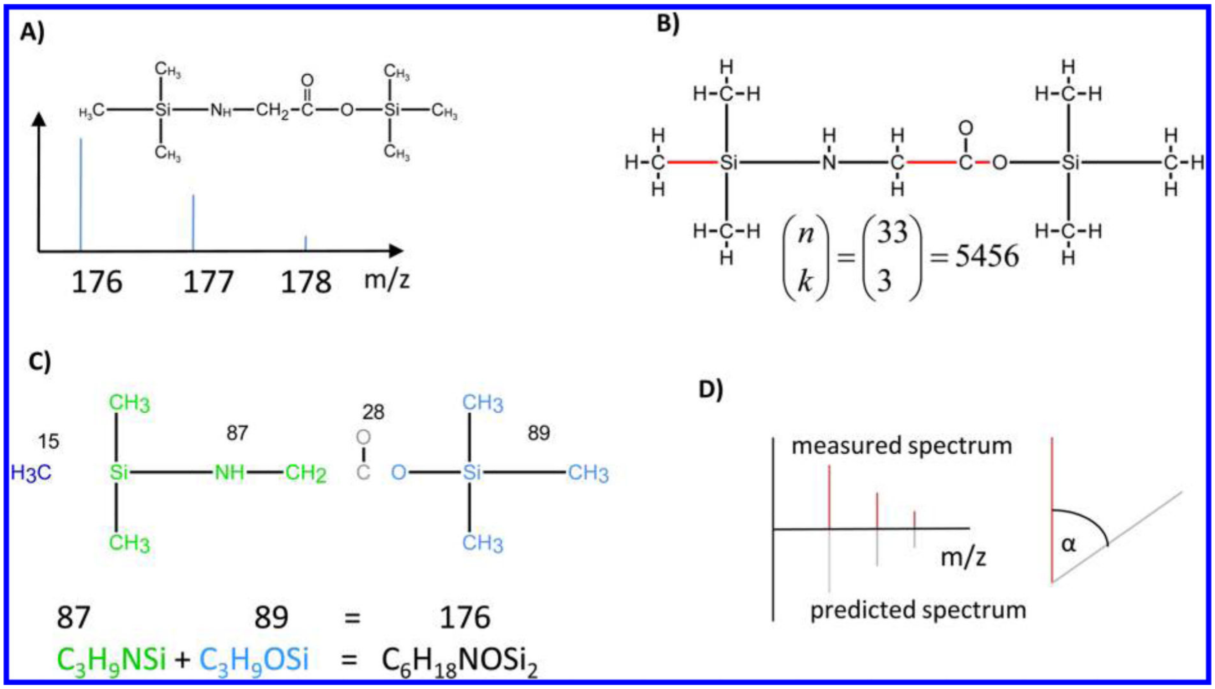

A)
H_3C—Si—NH—CH_2—C(=O)—O—Si—CH_3
176
177
178
m/z
B)
$\binom{n}{k} = \binom{33}{3} = 5456$
C)
15
H3C
CH3
Si
CH3
87
NH
CH2
28
O
C
O
CH3
Si
CH3
89
CH3
87 89 = 176
$C_3H_9NSi + C_3H_9OSi = C_6H_{18}NOSi_2$
D)
measured spectrum
m/z
predicted spectrum
α

Figure 2.

Overview of the algorithm. (A) As input FFC needs the 2D structure of the compound together with the mass spectrum of the ion of interest. In this example, we present the molecule *N*,*O*-bis-(trimethylsilyl)-glycine (219 Da) and the fragment ion at mass 176. (B) 2D Structure is first converted into a molecular graph. The graph contains 34 vertices and 33 edges. Then all combinations of edge sets of a certain size (in this case 3) are consecutively deleted from the graph, resulting in 5456 disconnected graphs, one for each edge set deleted. The number of resulting subgraphs can be calculated with the binomial coefficient, where *n* corresponds to the number of edges and *k* corresponds to the cut size (eq 3). For simplification, only the edge set leading to the correct fragmentation is shown here. (C) For each disconnected graph, the connected components are determined. For every combination of connected components where the molecular masses sum up to the mass of the fragment ion, the atoms of these components are combined to build up a candidate formula. In this example, the connected components shown in green and light blue with the masses 87 and 89 sum up to the target mass of 176. The candidate formula is then $C_6H_{18}NO_2Si_2$, which is indeed the correct formula for this fragment ion. In addition to the chemical formula, the algorithm also yields positional information about the fate of specific atoms. For example, the carboxyl carbon of the original glycine molecule is lost in this fragment ion. (D) On the basis of the candidate formula, the theoretical mass spectrum is predicted and a spectrum similarity score to the measured spectrum based on the dot product[17] is calculated. This is of special importance if more than one sum formula can be derived for the target mass.

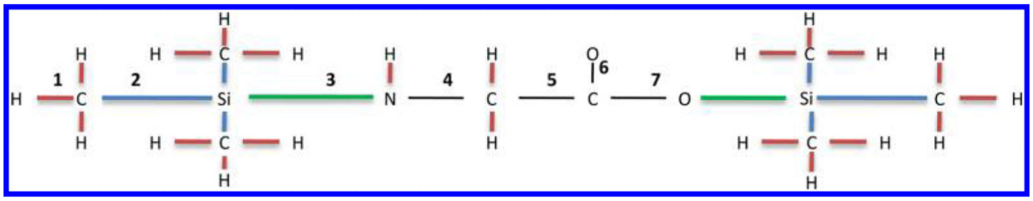


Figure 3.

Graph representation of *N*,*O*-bis-(trimethylsilyl)-glycine. The graph contains 33 edges. For a cut size of three, the number of distinct edge pairs to delete is 5456. To reduce the number of distinct edge pairs, non backbone edges (edges that are not connected to at least one backbone atom) are grouped based on their loss pattern. For example, edges shown in red are grouped together because their removal leads to the loss of one hydrogen. The group of edges shown in blue leads to the loss of a methyl group when one of these edges is removed. The group of edges shown in green lead to the loss of a TMS group when one of these edges is removed. After reduction to relevant backbone edges, the graph now contains only 7 distinct edge groups (as illustrated by the numbers above the edges) which reduces the number of distinct edge sets of size 3 from 5456 to 35.

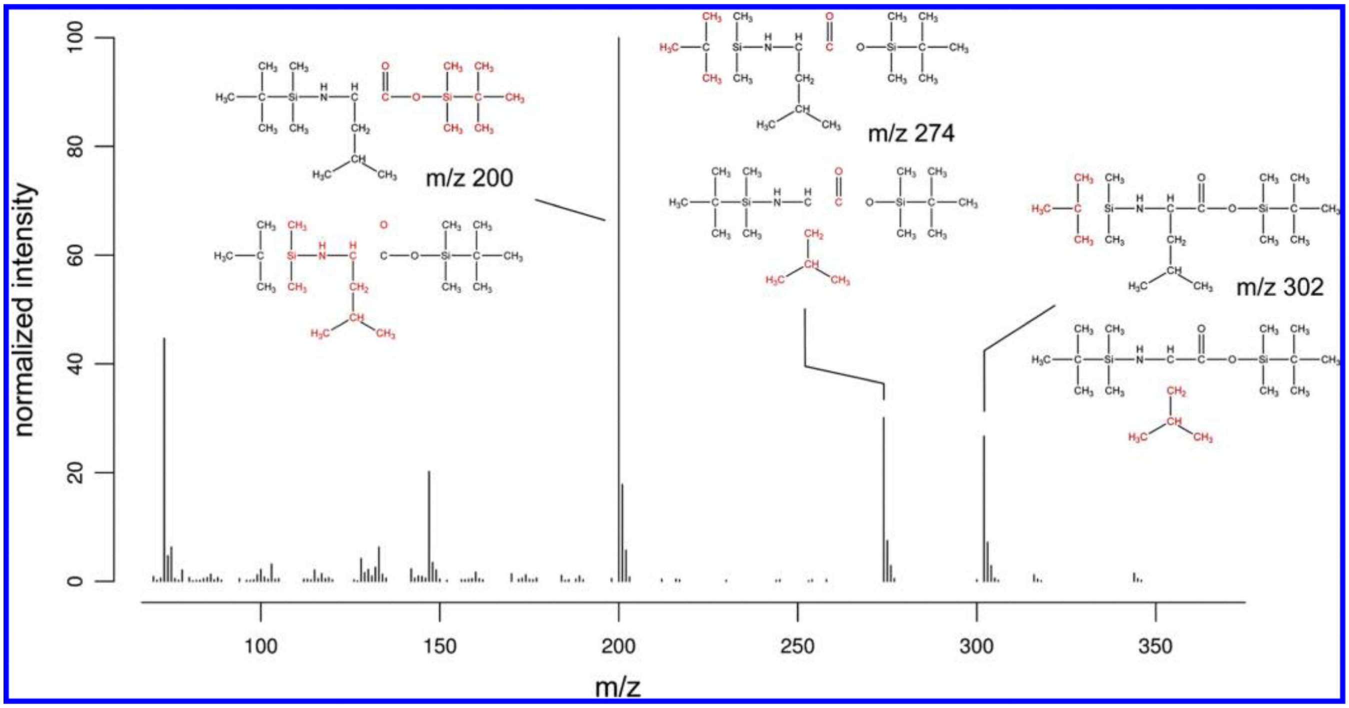


Figure 4.

Chemical formulas for ions 200, 274, and 302 of *N*,*O*-bis(dimethyl-*tert*-butylsilyl)-leucine. The two best-ranked hits, according to the number of broken bonds and spectrum similarity for each ion are shown. Incorrect fragmentations are visualized with a lower opacity, and cleaved atoms are shown in red.

Table 1

Fragments of TMS-Derivatized Compounds Part 1

compound	***m/z***	***m/z* ^{13}C**	***m/z* d_9-TMS**	**formula**
adenine 2TMS	279	284	297	$C_{11}H_{21}N_5Si_2$
	264	269	279	$C_{10}H_{18}N_5Si_2$
	206	211	215	$C_8H_{12}N_5Si$
alanine 2TMS	233	–	–	$C_9H_{23}NO_2Si_2$
	218	220, 221	233, 236	$C_8H_{20}NO_2Si_2$
	190	192	205	$C_7H_{20}NOSi_2$
	116	118	125	$C_5H_{14}NSi$
aspartic acid 2TMS	277	281	295	$C_{10}H_{23}NO_4Si_2$
	262	266	277	$C_9H_{20}NO_4Si_2$
	234	237	249	$C_8H_{20}NO_3Si_2$
	220	222	235	$C_7H_{18}NO_3Si_2$
	160	163	169	$C_6H_{14}N_1O_2Si$
aspartic acid 3TMS	349	354	376	$C_{13}H_{31}NO_4Si_3$
	334	338	358	$C_{12}H_{28}NO_4Si_3$
	306	309	330	$C_{11}H_{28}NO_3Si_3$
	292	294	316	$C_{10}H_{26}NO_3Si_3$
	232	235	250	$C_9H_{22}NO_2Si_2$
	218	220	236	$C_8H_{20}NO_2Si_2$
β-alanine 3TMS	305	–	–	$C_{12}H_{31}NO_2Si_3$
	290	–	314	$C_{11}H_{28}NO_2Si_3$
	248	–	272	$C_9H_{26}NOSi_3$
	232	–	250	$C_9H_{22}NO_2Si_2$
	174	–	192	$C_7H_{20}NSi$
	86	–	92	C_3H_6OSi
citric acid 4TMS	480	–	–	$C_{18}H_{40}O_7Si_4$
	465	471	498	$C_{17}H_{37}O_7Si_4$
	375	381	399	$C_{14}H_{27}O_6Si_3$
	363	368	390	$C_{14}H_{31}O_5Si_3$
	347	352	371	$C_{13}H_{27}O_5Si_3$
	273	278	291	$C_{11}H_{21}O_4Si_2$
3-phosphoglycerate 4TMS	474			$C_{15}H_{39}O_7PSi_4$
	459	462	492	$C_{14}H_{36}O_7PSi_4$
	387	387	423	$C_{12}H_{36}O_4PSi_4$
	357	359	384	$C_{11}H_{30}O_5PSi_3$
	315	315	342	$C_9H_{28}O_4PSi_3$
	299	299	323	$C_8H_{24}O_4PSi_3$

compound	***m/z***	***m/z*** **^{13}C**	***m/z d*$_9$-TMS**	**formula**
glycerol-3-phosphate 4TMS	460			$C_{15}H_{41}O_6PSi_4$
	445	448	478	$C_{14}H_{38}O_6PSi_4$
	387	387	423	$C_{12}H_{36}O_4PSi_4$
	357	359	384	$C_{11}H_{30}O_5PSi_3$
	341	343	365	$C_{10}H_{26}O_5PSi_3$
	299	299	323	$C_8H_{24}O_4PSi_3$

Table 2

Fragments of TMS-Derivatized Compounds Part 2

compound	*m/z*	*m/z* ^{13}C	*m/z* d_9-TMS	formula
glutamic acid 3TMS	363	368	390	$C_{14}H_{33}NO_4Si_3$
	348	353	372	$C_{13}H_{30}NO_4Si_3$
	320	324	344	$C_{12}H_{30}NO_3Si_3$
	246	250	264	$C_{10}H_{24}NO_2Si_2$
	230	234	245	$C_9H_{20}NO_2Si_2$
glutamine 3TMS	362	367	389	$C_{14}H_{34}N_2O_3Si_3$
	347	352	371	$C_{13}H_{31}N_2O_3Si_3$
	273	278	291	$C_{11}H_{25}N_2O_2Si_2$
	245	249	263	$C_{10}H_{25}N_2O_1Si_2$
glycerol 3TMS	308	–	–	$C_{12}H_{32}O_3Si_3$
	293	296	317	$C_{11}H_{29}O_3Si_3$
	218	221	236	$C_9H_{22}O_2Si_2$
	205	207	223	$C_8H_{21}O_2Si_2$
glycine 3TMS	291	293	–	$C_{11}H_{29}NO_2Si_3$
	276	278	300	$C_{10}H_{26}NO_2Si_3$
	248	249	274	$C_9H_{26}NOSi_3$
	174	175	192	$C_7H_{20}NSi_2$
isoleucine 2TMS	275	–	–	$C_{12}H_{29}NO_2Si_2$
	260	265, 266	275, 278	$C_{11}H_{26}NO_2Si_2$
	232	237	247	$C_{10}H_{26}NOSi_2$
	218	220	236	$C_8H_{20}NO_2Si_2$
	158	163	167	$C_8H_{20}NSi$
leucine 2TMS	275	–	–	$C_{12}H_{29}NO_2Si_2$
	260	265, 266	275, 278	$C_{11}H_{26}NO_2Si_2$
	232	237	247	$C_{10}H_{26}NOSi_2$
	218	220	236	$C_8H_{20}NO_2Si_2$
	158	163	167	$C_8H_{20}NSi$
lysine 3TMS	362	368	389	$C_{15}H_{38}N_2O_2Si_3$
	347	353	371	$C_{14}H_{35}N_2O_2Si_3$
	200	206	209	$C_9H_{18}NO_2Si$
	174	175	192	$C_7H_{20}NSi_2$
	156	161	165	$C_8H_{18}NSi$
lysine 4TMS	434	440	470	$C_{18}H_{46}N_2O_2Si_4$
	419	425	452	$C_{17}H_{43}N_2O_2Si_4$
	391	396	324	$C_{16}H_{43}N_2OSi_4$
	317	322	344	$C_{14}H_{37}N_2Si_3$

compound	***m/z***	***m/z* ^{13}C**	***m/z* d_9-TMS**	**formula**
	174	175	192	$C_{17}H_{20}NSi_2$
malic acid 3TMS	350	354	377	$C_{13}H_{30}O_5Si_3$
	335	339	359	$C_{12}H_{27}O_5Si_3$
	307	311	331	$C_{11}H_{27}NO_4Si_3$
	245	249	260	$C_9H_{17}O_4Si_2$
	233	236	251	$C_9H_{21}O_3Si_2$

Table 3

Fragments of TMS-Derivatized Compounds Part 3

compound	***m/z***	***m/z*** ^{13}**C**	***m/z*** d_9**-TMS**	**formula**
phenylalanine 2TMS	309	–	–	$C_{15}H_{27}NO_2Si_2$
	294	303	309	$C_{14}H_{24}NO_2Si_2$
	266	274	281	$C_{13}H_{24}NOSi_2$
	218	220	236	$C_8H_{20}NO_2Si_2$
	192	200	201	$C_{11}H_{18}NSi$
proline 2TMS	259	–	–	$C_{11}H_{25}NO_2Si_2$
	244	249	259	$C_{10}H_{22}NO_2Si_2$
	216	220	231	$C_9H_{22}NOSi_2$
	142	146	151	$C_7H_{16}NSi$
serine 3TMS	321	–	–	$C_{12}H_{31}NO_3Si_3$
	306	309	330	$C_{11}H_{28}NO_3Si_3$
	278	280	302	$C_{10}H_{28}NO_2Si_3$
	218	220	236	$C_8H_{20}NO_2Si_2$
	204	206	222	$C_8H_{22}NOSi_2$
	188	190	203	$C_7H_{18}NOSi_2$
succinic acid 2TMS	262		280	$C_{10}H_{22}O_4Si_2$
	247		262	$C_9H_{19}O_4Si_2$
	172		181	$C_7H_{12}O_3Si$
threonine 3TMS	335	–	–	$C_{13}H_{33}NO_3Si_3$
	320	324	344	$C_{12}H_{30}NO_3Si_3$
	218	221	236	$C_9H_{24}NOSi$
tyrosine 2TMS	325	–	–	$C_{15}H_{27}NO_3Si_2$
	310	319	325	$C_{14}H_{24}NO_3Si_2$
	282	290	297	$C_{13}H_{24}NO_2Si_2$
	208	216	217	$C_{11}H_{18}NOSi$
	192	200	198	$C_{10}H_{14}NOSi$
tyrosine 3TMS	397	–	–	$C_{18}H_{35}NO_3Si_3$
	382	391	406	$C_{17}H_{32}NO_3Si_3$
	354	362	378	$C_{16}H_{32}NO_2Si_3$
	280	288	298	$C_{14}H_{26}NOSi_2$
	218	220	236	$C_8H_{20}NO_2Si_2$
uracil 2TMS	256	260	284	$C_{10}H_{20}N_2O_2Si_2$
	241	245	256	$C_9H_{17}N_2O_2Si_2$
valine 2TMS	261	–	–	$C_{11}H_{27}NO_2Si_2$
	246	251	261	$C_{10}H_{24}NO_2Si_2$
	218	220, 222	233, 236	$C_9H_{24}NOSi_2$