

# Small Molecule LC-MS/MS Fragmentation Data Analysis and Application to Siderophore Identification

---

Oliver Baars and David H. Perlman

Additional information is available at the end of the chapter

<http://dx.doi.org/10.5772/63018>

---

### Abstract

Rapid developments in tandem liquid chromatography-mass spectrometry (LC-MS/MS) have created wide interest in applications for the analysis of small molecule mixtures. MS/MS spectra can contain rich structural information, but because of the structural diversity of small molecules and different data acquisition methods, analysis algorithms and workflows frequently need to be tailored to individual research questions. This chapter shows how MATLAB can be used for LC-MS/MS-based structural characterization of small molecules. Starting with the import of raw data, ways for visualization and the creation of graphical user interfaces (GUIs) for individual applications are demonstrated. A selection of frequently used algorithms for pre-processing and data analysis is reviewed in context of their MATLAB implementation. The approaches are then tailored and applied to the analysis of iron-binding peptides (peptidic siderophores) by high-resolution LC-MS/MS. The method uses a database with siderophore structures to exploit prior knowledge about siderophore structural diversity for the interpretation of MS/MS spectra from known and new siderophores.

**Keywords:** small molecules, metabolomics, fragmentation spectra, LC-MS/MS, liquid chromatography tandem mass spectrometry, neutral-loss, fragment-ion, auto-convolution spectra, molecular networks, secondary metabolites, siderophores, iron, nonribosomal peptides, MATLAB

---

### 1. Introduction

Liquid chromatography-mass spectrometry (LC-MS) enables the analysis of complex mixtures of small molecules and is applied widely in diverse research areas, such as metabolomics analysis in biology [1] and medicine [2, 3] and molecular characterization of samples in

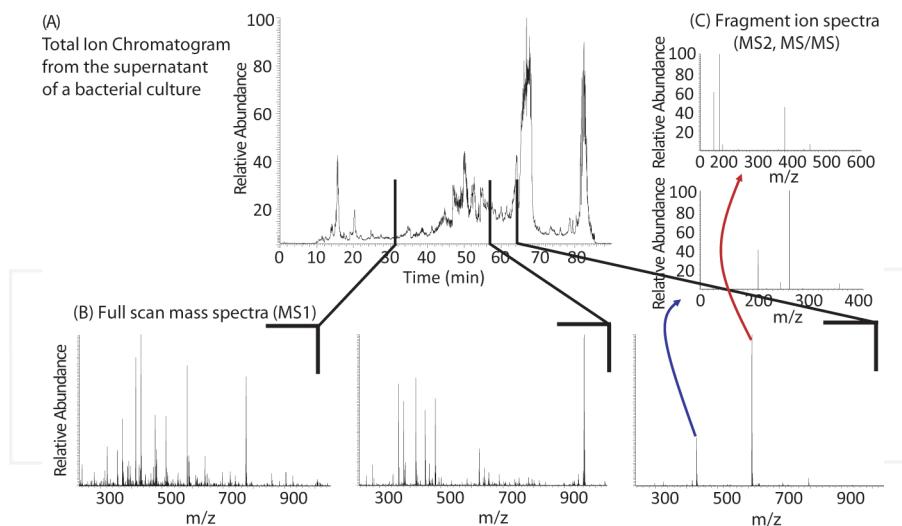
---

environmental chemistry [4] and combinatorial chemistry [5], among many other applications. Recent advances in LC-MS instrumentation with respect to speed and sensitivity, coupled with improved computational methods to extract information from complex datasets, have translated into falling costs of analyses and have created wide interest in LC-MS applications. As instruments have become able to explore samples at very high sensitivity and resolution (e.g., nano-flow LC coupled to high-resolution MS detectors, such as Orbitraps or Q-TOFs), the computational analysis of the generated raw data has become more challenging. In untargeted, ‘discovery type’ LC-MS experiments, the analysis of the raw data may include the following steps [6–8]:

1. Extraction of features from a raw LC-MS dataset. LC-MS features are defined by unique, characteristic combinations of retention time and mass-to-charge ratio. Associated with the chromatographic peak of a feature is a peak intensity or area, which serves as a relative measure of the abundance of the compound producing the feature. These parameters represent a fingerprint of compounds in a given sample.
2. Comparison of corresponding features in different sample sets and evaluation of significant differences.
3. Compound identification and structural characterization of unknown compounds of interest.

Soft ionization methods in LC-MS, most commonly electrospray ionization (ESI), generate mass spectra with minimal compound fragmentation and facilitate the extraction of LC-MS features associated with the intact molecular ion. Nevertheless, extraction of features for subsequent statistical analysis is non-trivial when complex mixtures of small molecules are analyzed [6]. Fortunately, open-source (e.g., mzMine2 [9], XCMS2 [10]) and commercial (e.g., Agilent MassHunter, Waters ProGenesis QI, ABSciex XCMSPlus, ThermoFisher Scientific SIEVE) software tools for this task have been developed and have become increasingly powerful and user-friendly.

With the extraction of features from LC-MS data becoming more readily achievable, the identification or structural characterization of small molecules has become a major bottleneck [11, 12]. Insight into chemical sum formulas and chemical interaction with the LC stationary phase (e.g., hydrophobicity in reversed phase chromatography) may be gained from LC-MS data because it contains information about molecular masses, isotope patterns, and retention times. However, even at ultra-high mass accuracies (<1 ppm), it is usually not possible to assign unique sum formulas to chromatographic features from MS1 data [13]. In addition, for any given sum formula, there are typically many theoretically possible isobaric compounds with different structures. Therefore, structural assignment of a compound is generally not possible based on MS1 data alone. Direct structural information can be obtained by measurement of fragmentation spectra (MS/MS, tandem MS, MS2, or for multiple rounds of fragmentation MS<sup>N</sup>). LC-MS/MS datasets are composed of time series of individual full scan MS1 spectra, interspersed by one or more MS/MS spectra derived from the fragmentation of one or more species present in the MS1 (**Figure 1**).



**Figure 1.** LC-MS/MS data is composed of time series of individual full scan MS1 spectra, interspersed by one or more MS/MS spectra. Shown are high-resolution LC-MS data for a bacterial culture supernatant collected with an Orbitrap XL mass analyser: (A) Total ion chromatogram (TIC, sum of MS1 intensities over time), (B) full scan MS1 spectra at three different retention times, and (C) fragment-ion spectra (MS2, MS/MS) for the two major peaks in the third MS1 spectrum.

Because of the vast diversity of small molecule structures, MS/MS-based structural characterization of compounds represents a great challenge [11, 12]. Identification by direct comparison of an experimental MS/MS spectrum to a library of MS/MS spectra is often limited by unavailable authentic standards. Even if MS/MS spectra for a compound (or a structurally closely related analogue) exist in a library, algorithms may not return the best match in the database, particularly if spectra are noisy or incomplete (e.g., contain contaminant ions, few fragments, minor fragments below detection limit, etc.) [14]. Complementary to these MS/MS spectral library-based approaches, de novo methods use known structures to calculate in-silico fragmentation spectra (random or rule based). Observed fragments are matched to possible substructures to reconstruct the observed MS/MS spectrum. The success of these methods depends on the chosen fragmentation rules and database search space [11]. Thus, the final success of attempts for structural characterization usually depends on a combination of prior knowledge about the compounds in the sample, adequate computational tools, and manual inspection of the raw data.

In addition to the identification of individual compounds, MS/MS structural information may be used at different stages in the interrogation of the samples. For example, at an early stage, it is possible to obtain an overview of structural diversity in a sample by clustering the MS/MS spectra into similarity networks [15], while at a later stage, once specific unknown features of interest have emerged, a more detailed interpretation of individual MS/MS spectra for compound characterization can be undertaken. MS/MS spectra can also be used to direct further targeted LC-MS/MS approaches aimed at deeper exploration of species possessing

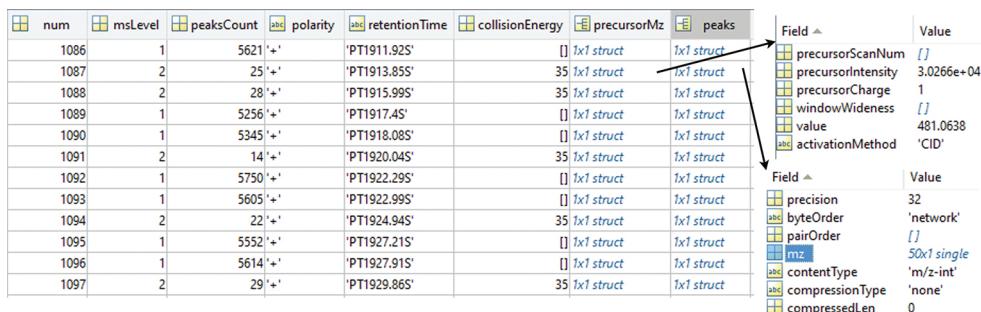
common fragments or fragmentation patterns of interest. This approach is commonly employed to specifically characterize molecules with certain functional groups or chemical modifications that produce characteristic patterns in their MS/MS spectra [14].

To make the best use of the rich structural information contained in LC-MS/MS datasets, there is a large need for MS/MS analysis algorithms that are tailored to many different individual research applications [11, 16]. As we demonstrate in this chapter, MATLAB provides an accessible and convenient platform for the interactive analysis and visualization of LC-MS/MS data and the implementation of customized algorithms and workflows. Tools are introduced for basic tasks, such as neutral-loss searches, as well as for more complex workflows, such as for the generation of MS/MS similarity networks or for the application of auto-convolution spectra to the structural characterization of peptides. We then apply these tools in the context of a specific basic research application: the discovery and structural characterization of peptidic siderophores. Siderophores are a class of secondary metabolites that are released by many bacteria and fungi to bind and take up iron (Fe), an essential and often growth-limiting micro-nutrient [17]. Using a siderophore structural database to exploit considerable prior knowledge about siderophore structural diversity, an effective workflow is shown for the LC-MS/MS-based analysis of known and new siderophores.

## 2. Importing raw data into MATLAB

The LC-MS raw data generated by the MS instrumentation is stored in vendor-specific binary file formats. For access by non-vendor software, the raw data file first needs to be converted to an open file format, such as the common mzXML format [7]. The program msconvert, which is part of the open-source proteomics package ProteoWizard (<http://proteowizard.sourceforge.net/tools.shtml>), can be used to convert from most raw data formats to mzXML [18]. At this point, LC-MS data recorded in profile mode can also be centroided by vendor supported algorithms built into msconvert. During centroiding, the centroid is determined for each mass spectral peak, which consists of multiple  $m/z$ —intensity measurements across the profile of the detected ion at any given time, and the peak is replaced by a single  $m/z$  and intensity pair at its center [18]. Centroiding reduces the data file size and is required for the subsequent analysis steps in this chapter.

MATLAB's bioinformatics toolbox provides several functions applicable to the processing of LC-MS data. For this study, we will use the functions mzxmlread and mzxml2peaks. The function mzxmlread can be used to import LC-MS data from mzXML files into a MATLAB structure array: `mzXMLstruct = mzxmlread(mzXMLFilename)`. The returned structure contains the LC-MS/MS data and relevant metadata from the mzXML file, such as scan number, MS level, ionization mode, MS/MS collision energy, and MS/MS precursor information (**Figure 2**).



The figure shows a MATLAB Data Browser window. On the left is a table with columns: num, msLevel, peaksCount, polarity, retentionTime, collisionEnergy, precursorMz, and peaks. The 'peaks' column contains 1x1 struct arrays. On the right, a detailed view of the 'peaks' field is shown, listing its properties: precursorScanNum, precursorIntensity, precursorCharge, windowWideness, value, activationMethod, precision, byteOrder, pairOrder, m/z, contentType, compressionType, and compressedLen. Arrows point from the table to the detailed view.

num	msLevel	peaksCount	polarity	retentionTime	collisionEnergy	precursorMz	peaks
1086	1	5621 <sup>+</sup>	'PT1911.92S'		[] 1x1 struct	[] 1x1 struct	
1087	2	25 <sup>+</sup>	'PT1913.85S'	35	1x1 struct	1x1 struct	
1088	2	28 <sup>+</sup>	'PT1915.99S'	35	1x1 struct	1x1 struct	
1089	1	5256 <sup>+</sup>	'PT1917.4S'	[] 1x1 struct	[] 1x1 struct		
1090	1	5345 <sup>+</sup>	'PT1918.08S'	[] 1x1 struct	[] 1x1 struct		
1091	2	14 <sup>+</sup>	'PT1920.04S'	35	1x1 struct	1x1 struct	
1092	1	5750 <sup>+</sup>	'PT1922.29S'	[] 1x1 struct	[] 1x1 struct		
1093	1	5605 <sup>+</sup>	'PT1922.99S'	[] 1x1 struct	[] 1x1 struct		
1094	2	22 <sup>+</sup>	'PT1924.94S'	35	1x1 struct	1x1 struct	
1095	1	5552 <sup>+</sup>	'PT1927.21S'	[] 1x1 struct	[] 1x1 struct		
1096	1	5614 <sup>+</sup>	'PT1927.91S'	[] 1x1 struct	[] 1x1 struct		
1097	2	29 <sup>+</sup>	'PT1929.86S'	35	1x1 struct	1x1 struct	

Field ▲ Value  
 precursorScanNum 1  
 precursorIntensity 3.0266e+04  
 precursorCharge 1  
 windowWideness []  
 value 481.0638  
 activationMethod 'CID'  
  
 Field ▲ Value  
 precision 32  
 byteOrder 'network'  
 pairOrder []  
 m/z 50x single  
 contentType 'm/z-int'  
 compressionType 'none'  
 compressedLen 0

Figure 2. A selection of fields in the MATLAB structure array returned after the import of an mzXML file with LC-MS/MS data.

The function `mzxml2peaks` can be employed to extract the mass spectra (*m/z* values and intensities), together with their corresponding retention times, for a selected MS level (MS1, MS2, etc.) from `mzXMLstruct`: `[Spectra, Times] = mzxml2peaks(mzXMLstruct, MSLevel, LevelValue)`, where Spectra is a cell array in which each element contains a two-column matrix of *m/z*—intensity pairs corresponding to one mass spectrum collected at a specific retention time, and Times is a vector with the corresponding retention time for each mass spectrum. To analyze MS/MS spectra, it is usually necessary to retrieve additional information from `mzXMLstruct`. Information about MS/MS precursor ions (precursor *m/z*, intensity, and charge) can be accessed in `mzXMLstruct.scan.precursorMz`:

```
% Number of peaks (Peak Count) for each spectrum
PC = arrayfun(@(x) x.peaksCount, mzXMLstruct.scan);
% MS level of each spectrum
level = arrayfun(@(x) x.msLevel, mzXMLstruct.scan);
h = find(PC & ismember(level,2)); % select only MS2 spectra
% extract precursor m/z, intensity, and charge
prec_mz = arrayfun(@(x) x.precursorMz.value, mzXMLstruct.scan(h));
prec_Int = arrayfun(@(x) x.precursorMz.precursorIntensity, ...
  mzXMLstruct.scan(h));
prec_z = arrayfun(@(x) x.precursorMz.precursorCharge, mzXMLstruct.scan(h));
```

If mass spectra are collected in positive and negative ionization modes during the same run, it may be useful to process positive and negative modes separately. To extract one ionization mode only, a filter can be applied to the data in `mzXMLstruct`:

```
Pol = '+'; % only positive mode data
pol_select = arrayfun(@(x) strcmp(x.polarity, Pol), mzXMLstruct.scan);
mzXMLstruct.scan = mzXMLstruct.scan(pol_select);
```

These loops through scan metadata in mzXMLstruct can be added to the mzxml2peaks function together with optional output and input arguments. In the following, we will use an optional output matrix called Precursors which contains precursor information for each MS/MS scan in three columns:  $m/z$ , intensity, and  $z$ .

### 3. Visualization and graphical-user-interface implementation

When analyzing LC-MS/MS data, it is helpful to be able to visualize spectra and chromatograms in order to display significant features, evaluate spectral noise, consider potential interferences, or evaluate fragmentation patterns, and so on. The two common projections of the three-dimensional LC-MS data (retention time,  $m/z$ , intensity) into two-dimensional space are chromatograms (intensity vs. retention time) and mass spectra (intensity vs.  $m/z$ ). Note that a chromatogram with intensities of a selected range of  $m/z$  values is called an extracted ion chromatogram (EIC), whereas a chromatogram that shows the sum of all ions is known as a total ion chromatogram (TIC). The  $m/z$  tolerance  $\Delta m/z$  is often given in ppm of the  $m/z$  value or as an absolute error in amu.

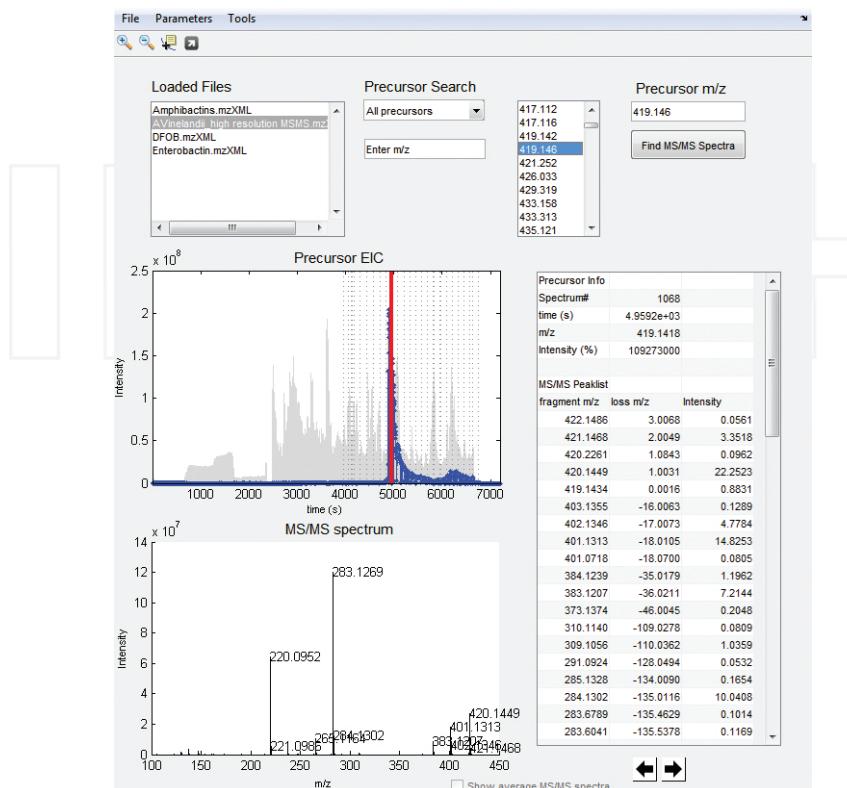
To plot an MS/MS spectrum for a given precursor  $m/z$ , the matrix Precursors can be used to find precursor  $m/z$  values within a specified error tolerance. The corresponding elements in Spectra represent the MS/MS spectra recorded for this  $m/z$  value:

```
mz = 419.142; % find MS/MS spectra for precursor with m/z = 419.142
% ppm error or minimum absolute error
error = max(minerror, ppmerror * 10^-6 * mz);
MS2idx = find(abs(Precursors(:,1) - mz) <= error);
if ~isempty(MS2idx)
    MS2Peaks = Spectra{MS2idx(1)}; % select the first found MS2 spectrum
end

% To plot the centroided mass spectrum:
plotms(2:3:3 * size(MS2Peaks,1),2) = MS2Peaks (:,2);
plotms(1:3:3 * size(MS2Peaks,1),1) = MS2Peaks (:,1);
plotms(2:3:3 * size(MS2Peaks,1),1) = MS2Peaks (:,1);
plotms(3:3:3 * size(MS2Peaks,1),1) = MS2Peaks (:,1);
plot(plotms(:,1), plotms(:,2));
```

Graphical-user-interfaces (GUIs) can aid in efficiently browsing and evaluating the LC-MS/MS data and can serve as a platform for data manipulation and analysis. GUIs are also particularly helpful for users who are not familiar with MATLAB and can be shared as stand-alone applications. MATLAB facilitates the creation and programming of GUIs with the tool GUIDE (graphical user interface design environment). GUIDE provides an interface for designing the layout, while creating the code for the GUI, including the implementation of callbacks for a number of standard User Interface (UI) controls, such as 'pushbuttons', 'popup menus', and 'listboxes'. The callback functions can then be filled with user-defined code to

provide the intended functionality. A GUI for LC-MS/MS data analysis created with MATLAB is shown in **Figure 3**.



**Figure 3.** Graphical User Interface (GUI) for visualization, pre-processing, and analysis of LC-MS/MS data. The main window allows the user to select or search MS/MS precursor  $m/z$  values. When a precursor  $m/z$  value is selected, a corresponding EIC is generated with the MS1 data and the MS/MS spectrum is shown in this figure. The retention time corresponding to the MS/MS spectrum is indicated by a red line in the EIC figure. A table to the right shows the precursor information and a table of the MS/MS peaks. A menu bar provides functionality to import LC-MS data from mzXML files, open or save files, enter  $m/z$  error tolerances and other parameters, and select from a range of pre-processing and analysis tools.

#### 4. Pre-processing of MS/MS spectra

Before analysis of LC-MS/MS data, pre-processing can be employed to increase signal-to-noise ratios, remove contaminant peaks, and reduce the data for the following analysis steps. Pre-processing and analysis algorithms are often tailored to individual data acquisition methods, data quality, and analytical goals.

#### 4.1. Noise removal

The ratio of maximum-to-median peak intensities in MS/MS spectra has been used as an estimate of signal-to-noise ratios [14]. In order to remove noise and reduce the amount of data in MS/MS spectra, filters have been applied to retain only the most intense peaks within a given mass spectral bracket (e.g., five most intense peaks in a 50 Da window) [19, 20]. If several fragment-ion spectra are available for the same precursor, noise can be removed by retaining only those fragments that are present in a majority of the MS/MS spectra [21] or the spectra can be summed or averaged, which may minimize the noise contribution if it is random (see Section 4.3). For high-resolution MS/MS spectra, exact masses can also be used to determine possible fragment-ion sum formulas in order to remove noise or satellite peaks that possess *m/z* values leading to possible compositions that are completely incongruent with the precursor [21]. Similarly, MS/MS peaks with masses larger than the precursor mass can be removed.

Interfering signals are not only due to random noise but frequently a consequence of relatively wide precursor ion isolation windows (usually > 1 Da) utilized by the instruments' ion optics during ion selection prior to MS/MS. For this reason, the selection of minor features that may be well resolved in MS1 can result in the co-isolation of significant quantities of unrelated species, which, in turn, produce significant contaminant peaks in the corresponding MS/MS spectra. Spectral deconvolution algorithms to remove possible contaminant peaks or to determine multiple precursors from unintentional or intentional wide-window ion isolation (such as that achieved in recently developed data-independent acquisition methods) have been published [22, 23].

#### 4.2. Removal of $^{13}\text{C}$ -isotopologue peaks

Another consequence of wide precursor isolation windows is that fragment ions may be accompanied to some degree by their  $^{13}\text{C}$ -isotopologues. To de-isotope a high-resolution MS/MS spectrum, an algorithm can proceed from the lowest to the highest intensity fragment ion. For each ion peak, it is evaluated whether it may represent a  $^{13}\text{C}$  isotopologue of a more abundant peak, by searching for another peak with the exact mass difference of a  $^{13}\text{C}$  isotope ( $\Delta m/z = -1.0034 \text{ Da}$  for  $z = 1$ ), within the defined error tolerance (e.g.,  $\pm 0.0035 \text{ Da}$ ). If this mass difference is detected, the  $^{13}\text{C}$  isotopologue peak is removed from the spectrum. This process is repeated for all relevant charge states, for example, all charge states up to the precursor charge. At the same time, if  $^{13}\text{C}$  isotopologue peaks are detected, this procedure can be used to assign charge states to individual fragment ions.

#### 4.3. Consensus spectra

If several MS/MS spectra of the same precursor have been acquired, the construction of consensus spectra can increase signal-to-noise ratios and significantly speed up the downstream analysis. To select which MS/MS spectra are to contribute to one consensus spectrum, the Spectra cell array can be filtered by identifying the elements with the same precursor *m/z* (within the specified *m/z* error tolerance) and charge state *z* in the Precursor matrix. In many situations, a number of different possible precursor structures can exist for the same *m/z* value

(isobaric compounds) or indistinguishable  $m/z$  values within specified  $m/z$  tolerances. In such cases, the retention time can be used as an additional filter, that is, only those MS/MS spectra are clustered that are recorded during elution of the corresponding parent ion in MS1. It is also possible to combine only those MS/MS spectra in a consensus spectrum that show high pairwise similarity [20, 21]. The calculation of MS/MS spectral similarity is described in detail in Section 5.2.

To calculate a consensus spectrum for each cluster of MS/MS spectra, all  $m/z$ -intensity pairs of the individual contributing MS/MS spectra are combined in a common matrix [21]. All peaks within the defined  $m/z$  tolerance around the most intense peak are determined and an intensity-weighted mean  $m/z$ , and an intensity value can be calculated and then refined iteratively, as follows. If any additional peaks are within the bin around the intensity-weighted average  $m/z$ , a new intensity-weighted average is calculated until no further peaks fall within the bin. The peaks within the bin are then replaced by the intensity-weighted average  $m/z$  and intensity value and the process is repeated with all peaks in order of decreasing peak intensity.

## 5. Analysis of fragmentation spectra

In this section, three relevant LC-MS/MS analysis approaches are reviewed that can be applied for a wide range of analytical questions. In Section 6, we illustrate an example application for each approach (Sections 6.2–6.4) that is specifically tailored to the discovery and structural characterization of siderophores.

### 5.1. Fragment-ion and neutral-loss search

Fragment-ion or neutral-loss searches can be utilized to mine LC-MS/MS data for molecules with characteristic substructures or fragmentation behavior, such as certain lipid headgroups [24] or metabolite conjugates (e.g., GSH or phosphate) [25]. In targeted analysis, a defined fragment-ion or neutral-loss can be exploited to increase specificity for detection of the target compound.

Simple algorithms can loop through the MS/MS cell array (Spectra) to search for fragment-ion peaks within the defined  $m/z$  tolerance. To detect defined neutral-losses, ‘neutral-loss peaks’ can be computed as differences between the  $m/z$  values of the precursor ion and the fragment-ions (precursor neutral-loss search) or all pairwise differences between the measured ions (including the precursor and all fragment-ions). The algorithm can be refined by taking into account the charge state of the precursor ion and the fragment-ions to calculate fragment-ion and neutral-loss masses instead of  $m/z$  values (see Section 4.2 for fragment-ion charge state determination).

### 5.2. Pairwise similarity of MS/MS spectra

Given the task to find a best match for an experimental MS/MS spectrum among members of a spectral library, the experimental spectrum needs to be compared to each spectrum in the

database and a pairwise similarity score needs to be computed. A common approach is to calculate a normalized dot product (cosine similarity) between pairs of MS/MS spectra  $S_A$  and  $S_B$  [14]:

$$\text{cosine similarity score} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}} \quad (1)$$

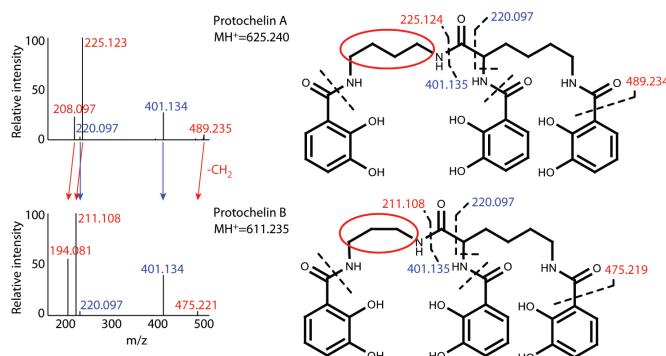
**A** and **B** are two vectors that contain the peak intensities from the MS/MS spectra  $S_A$  or  $S_B$ . The intensities of two peaks that occur at the same  $m/z$  in  $S_A$  and  $S_B$  (within the defined  $m/z$  tolerance) are corresponding elements  $A_i$  and  $B_i$ . To match fragment-ions in  $S_A$  and  $S_B$ , the algorithm can proceed from high to low intensities and identify corresponding peaks within a defined absolute or relative  $m/z$  tolerance in the corresponding other spectrum [20, 21]. If a peak is present in only one of the two spectra, its intensity is added to the respective vector **A** or **B** while the corresponding element in the other vector is set to 0. A similarity score of 1 indicates identical spectra, whereas a value of 0 indicates that no fragments with a common  $m/z$  are present. An implementation of this approach in MATLAB is as follows:

If a matching structure is not present in the database, the information may nevertheless be used for the identification of potential common substructures of a structurally related com-

```
% calculate normalized intensities for each MS2 spectrum
Spectra_norm = arrayfun(@(x) [x{1}(:,1) x{1}(:,2) ./ max(x{1}(:,2))], ...
    Spectra, 'UniformOutput', false);
error = 0.005; % m/z tolerance
% select the first two MS2 spectra for calculation of cosine similarity
MS2_A = sortrows(Spectra_norm{1}, 2); % sort with ascending peak intensity
MS2_B = sortrows(Spectra_norm{2}, 2);
% matched_AB will contain intensities of matched peaks from MS2_A and MS2_B
% in columns 1 and 2
matched_AB = [MS2_A(:,2) zeros(size(MS2_A,1),1)];
for i = size(MS2_A,1):-1:1 % in order of decreasing peak intensity in MS_A
    if ~isempty(MS2_B)
        common = abs(MS2_B(:,1) - MS2_A(i,1)) <= error; % matching m/z in MS_B?
        common = common .* MS2_B(:,2); % corresponding intensities in MS_B
        % if more than 1 match is found, select the one with highest intensity
        if sum(common) > 0
            maxB_row = find(common, 1, 'last');
            matched_AB(i,2) = MS2_B(maxB_row,2);
            MS2_B(maxB_row,:) = []; % delete matched peak in MS2_B
        end
    end
end
% add unmatched peaks from MS2_B to matched_AB
matched_AB = [matched_AB; [zeros(size(MS2_B,1),1) MS2_B(:,2)]];
A = matched_AB(:,1); B = matched_AB(:,2);
similarity_AB = dot(A,B) / (norm(A) * norm(B)); % cosine similarity
```

ound. For such applications, the matching of fragments in **A** and **B** can be modified to include not only common fragment-ions in both spectra but also common neutral-losses, that is, pairs of peaks in  $S_A$  and  $S_B$  that have  $m/z$  values which differ by a common  $m/z$  value. In a simple implementation this difference can be the mass difference of the parent ions of  $S_A$  and  $S_B$  (see **Figure 4** for an example) [25, 26].

Another useful application of similarity calculations is found in the generation of MS/MS similarity networks as described previously [15]. Calculating all pairwise similarities between consensus LC-MS/MS spectra yields a table that can be visualized in a molecular network using freeware tools such as Cytoscape ([www.cytoscape.org](http://www.cytoscape.org)). In the MS/MS network, each node represents one consensus spectrum (precursor information) and each edge between two nodes illustrates the relatedness. Cytoscape provides functionality to create edge-weighted force-directed layouts to cluster closely related nodes in order to obtain an overview of structural diversity in a sample.



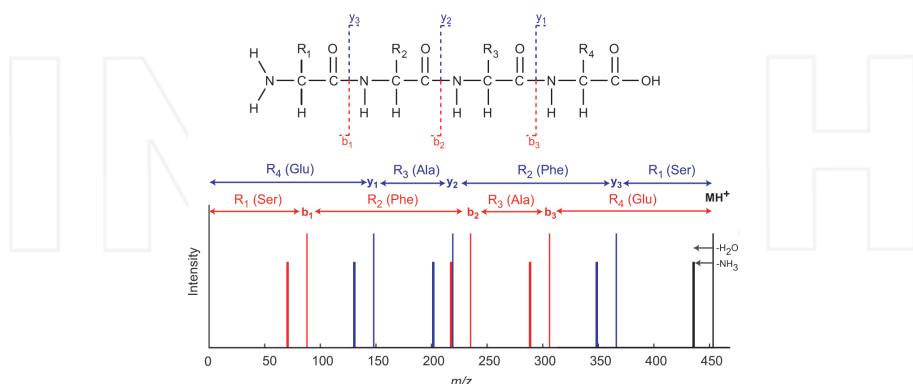
**Figure 4.** MS/MS spectra of the siderophore protocochelin A and its related analogue protocochelin B. Protocochelin A and B have the same structures except for an exchange of a 1,4-diaminobutane linker group in protocochelin A with 1,3-diaminopropane in protocochelin B (red circles). Accordingly, parent ion  $m/z$  ratios differ by a  $\text{CH}_2$  group and fragment peaks that include the modification are shifted by the  $m/z$  of  $\text{CH}_2$  ( $\Delta m/z = 14.0157$ ). To calculate similarities between such structurally related compounds, the shifted peaks can be matched in addition to the peaks that both spectra have in common. Figure modified from [27].

### 5.3. MS/MS convolution and auto-convolution

De-novo sequencing of peptides is most widely performed by the analysis of fragmentation spectra that are acquired by positive-mode collision-induced dissociation (CID) [28]. With positive-mode CID, major MS/MS peaks result from dissociation of the molecule at the peptide bonds, yielding b- and y-type ions (**Figure 5**) [29, 30]. In addition, the spectra regularly include other related fragments. For example a-ions have a mass difference corresponding to a CO group relative to b-ions ( $\Delta m = 27.9949$ ) and, if present, can be used to distinguish b-ions from y-ions. In addition, neutral-losses of  $\text{H}_2\text{O}$  or  $\text{NH}_3$  are often observed.

As illustrated in **Figure 5**, the mass differences between fragments include the mass of individual peptide monomers (amino acid residues). Spectral convolution between two spectra  $S_A$  and  $S_B$  calculates the  $m/z$  difference between each peak in  $S_A$  and each peak in  $S_B$ . The multiplicity of each observed  $m/z$  difference, within the given  $m/z$  tolerance, is then counted to yield the convolution spectrum with the multiplicity for each  $m/z$  versus the observed  $m/z$  differences. The convolution spectrum between unrelated spectra  $S_A$  and  $S_B$  is close to 0 or 1 for most  $m/z$  values, whereas the convolution spectrum for structurally related peptides (i.e., peptides with shared sequences) will show significant peaks for some  $m/z$  values [26]. The following MATLAB code shows a basic implementation of spectral convolution.

```
% select two MS2 spectra for convolution and get the m/z of all peaks
mz_A = Spectra{1}(:,1); mz_B = Spectra{2}(:,1);
error = 0.005; % m/z tolerance
PC_A = numel(mz_A); PC_B = numel(mz_B); n_AB = PC_A * PC_B;
diffAB = zeros(n_AB,1); % diffAB will be a vector with all m/z differences
MS2_conv = nan(n_AB,2); % MS2_conv will be the convolution spectrum
for i = 1:PC_B
    diffAB((i - 1) * PC_A + 1:i * PC_A) = abs(mz_B(i,1) - mz_A(:,1));
end
n = 0;
% group m/z within m/z tolerance in diffAB
while ~isempty(diffAB)
    n = n + 1;
    bin_idx = abs(diffAB - diffAB(1)) <= error;
    % mean of m/z differences and multiplicity
    MS2_conv(n,:) = [mean(diffAB(bin_idx)) sum(bin_idx)];
    diffAB(bin_idx) = []; % delete processed m/z values in diffAB
end
MS2_conv(isnan(MS2_conv(:,1)),:) = [];
```



**Figure 5.** Theoretical LC-MS spectrum for a peptide with the sequence Ser-Phe-Ala-Glu. The spectrum shows the position of expected b- and y-ions together with neutral-losses of  $\text{H}_2\text{O}$  or  $\text{NH}_3$ .

Auto-convolution spectra are generated by calculating the *m/z* differences between all peaks within one single spectrum. These have been used for the identification of possible peptide monomers in cyclic nonribosomal peptides (NRPs) [19]. NRPs are secondary metabolite peptides synthesized by nonribosomal peptide synthetases (NRPS), and include antibiotics, toxins, and siderophores. The structures of NRPs contain unusual non-proteinogenic amino acids, which increase the number of possible monomers from the canonical 20 found in most proteins to several hundred. Possible peptide monomer species in cyclic NRPs are revealed by matching peaks in the auto-convolution spectrum to masses in a database of possible peptide monomers [19]. The NORINE database with NRPs and corresponding peptide monomers can be used for this purpose and is freely accessible at <http://bioinfo.lifl.fr/norine/> [31].

## 6. Application to siderophore analysis

In this section, the LC-MS/MS analysis methods discussed above are applied to the discovery and structural characterization of peptidic siderophores. Siderophores are secondary metabolites that are released by many bacteria and fungi to bind and take up iron (Fe), an essential and often growth-limiting micro-nutrient [17]. Using a siderophore structural database to exploit considerable prior knowledge about siderophore structural diversity, an effective workflow is presented for the LC-MS/MS-based analysis of known and new siderophores.

### 6.1. Workflow

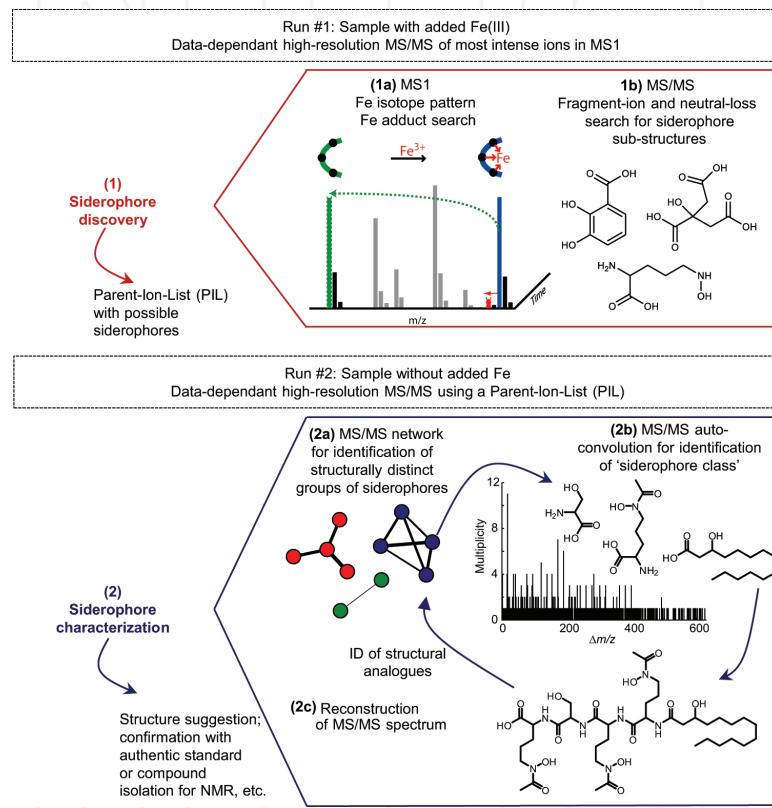
#### 6.1.1. Overview

Previously, we described an algorithm for the discovery of siderophores in high-resolution LC-MS1 data by screening for the natural stable isotope pattern of iron ( $^{54}\text{Fe}$  and  $^{56}\text{Fe}$ ) bound to siderophores and by searching for related iron-free siderophores [32]. Here, we complement this method by analysis of high-resolution LC-MS/MS data for discovery and structural characterization of siderophores (**Figure 6**).

To obtain a list of siderophore candidates, Fe can be added to the sample extract before injection onto the LC-MS system, which facilitates the generation of Fe-ligand complexes and the recognition of the Fe isotope patterns associated with Fe-bound siderophores (**Figure 6-1a**). Independent of isotope patterns, fragmentation spectra can be mined for siderophore-characteristic substructures by fragment-ion and neutral-loss searches (**Figure 6-1b**, Section 6.2). Both approaches yield a table with *m/z* ratios of candidate siderophore Fe complexes and associated free siderophores. The tables are combined to create a parent-ion-list for a replicate run with data-dependent LC-MS/MS acquisition.

For the replicate run, no Fe is added to the sample extract, maximizing the signal for the free siderophore species, which are preferentially selected for structural characterization in subsequent analytical steps (at the same time, differences in peak abundances between the extracts with and without added Fe can give further confidence to the assignment of sidero-

phores). MS/MS spectra of unbound siderophore candidates are selected to generate an MS/MS molecular network which provides an overview of structurally distinct groups of siderophores (**Figure 6-2a**, Section 6.3). Representative species in the network are selected for structural characterization by calculation of auto-convolution spectra (**Figure 6-2b**, Section 6.4). By matching peaks in the auto-convolution spectra to masses in a database of siderophore peptide monomers, possible siderophore substructures are assigned. Combinations of peptide-monomers are then used as a signature to find possible related known structures in the database, aiding in the reconstruction of the original MS/MS spectrum (**Figure 6-2c**). Iterations with structurally related compounds in the MS/MS network can refine the structure suggestions and be used to efficiently evaluate structures of derivatives. Depending on the



**Figure 6.** Schematic workflow for discovery (6-1) and structural characterization (6-2) of siderophores by high-resolution LC-MS/MS. Two complementary approaches may be used for siderophore discovery: mining for characteristic iron-isotopic patterns of iron siderophore complexes as well as for peaks corresponding to unbound siderophore species (6-1a), and searches for fragment-ions or neutral-losses associated with characteristic siderophore substructures (6-1b). To characterize siderophore structures, MS/MS similarity networks (6-2a) can be used to identify groups of structurally distinct siderophores. MS/MS auto-convolution in combination with a siderophore database may reveal sub-structures and ‘siderophore class’ (6-2b) to aid in the reconstruction of the original MS/MS spectrum (6-2c). The analysis results may then inform iterations with structurally related siderophores in the MS/MS network.

analysis outcomes, the putative structures can be confirmed with authentic standards or by isolation of the compound for characterization by orthogonal means (e.g., by nuclear magnetic resonance spectroscopy, etc.).

#### 6.1.2. Experimental methods and data pre-processing

The samples for this study include the siderophore standards desferrioxamine B (DFOB, Aldrich), enterobactin (EMC Biochemicals), amphibactins (kindly provided by A. Butler, UC Santa Barbara), and extracts of iron-limited *Azotobacter vinelandii* culture supernatants. *A. vinelandii* culture conditions and sample preparation were described previously [27].

LC-MS analyses were performed on a high-performance liquid chromatography (HPLC)-MS platform, using a C18 column coupled to an LTQ-Orbitrap XL hybrid mass spectrometer (ThermoFisher). Samples were separated under a gradient of solutions A and B (solution A consisted of water, 0.1% FA, and 0.1% acetic acid; solution B consisted of acetonitrile, 0.1% FA, and 0.1% acetic acid; gradient, 0 to 100% B; flow rate, 50  $\mu$ l/min). Full-scan mass spectra were acquired in positive-ion mode with a resolving power (R) of 60,000 ( $m/z$  = 400). MS/MS spectra were simultaneously acquired using collision-induced dissociation (CID; 35 V collision voltage) in the Orbitrap, a parent ion intensity threshold of 10,000, and targeting the three most abundant species in the full-scan spectrum or targeting selectively only predefined species on a parent ion list.

LC-MS/MS raw data are converted to mzXML, centred, and imported into MATLAB as described in Section 2. Spectra in which the maximum-to-median intensity ratio is below 3 are removed. Spectra are de-isotoped with an  $m/z$  tolerance for  $^{13}\text{C}$  isotopes of  $\Delta m/z$  = 0.0035, and only the top five most intense peaks in a 50 Da window are retained. The user-defined  $m/z$  tolerance for neutral-loss and parent-ion searches is set to  $\Delta m/z$  = 0.0050. For the generation of MS/MS molecular networks and auto-convolution spectra, consensus spectra are calculated using an  $m/z$  bin width of 0.01 Da ( $\Delta m/z$  = 0.0050). A siderophore database with >300 known siderophore structures was assembled in ChemBioFinder<sup>TM</sup> to determine siderophore-characteristic substructures, and to aid in the structural characterization of known and new siderophores.

## 6.2. Fragment-ion and neutral-loss searches

Utilizing a database with >300 known siderophore structures, the most frequently occurring iron binding substructures in siderophores are identified (**Table 1, Figure 7**). The specificity of these structures for siderophore discovery by fragment-ion or neutral-loss searches is evaluated by searching the NORINE database of nonribosomal peptides (NRPs) with >1,100 NRP structures (<http://bioinfo.lifl.fr/norine/>). With the exception of N-hydroxyornithine and N-cyclo-hydroxyornithine, the peptide monomers have unique masses in NORINE ( $\pm 0.005$  Da, Table 1). In addition, the iron binding substructures are not observed in non-siderophore structures in the database, with the exception of two putative NRPS products that are predicted to include compounds 10 and 11. Corresponding neutral-loss searches in the METLIN library with >70,000 small molecule CID MS/MS spectra reveal 39–232 false positive neutral-loss hits (i.e., matching neutral-loss mass within  $\pm 0.005$  Da despite siderophore-unrelated structures), representing less than 1% of the MS/MS spectra in the database.

#	Monomer name	Monoisotope mass	Neutral loss mass	Unrelated isobaric structures in NORINE(*)	Occurrence in non-siderophore structures in NORINE (**)	False-positive neutral-loss hits in METLIN (***)
1	N-acetyl-hydroxy-ornithine (Ac-OH-Orn)	190.0954	172.0848	0	0	122
2	N-formyl-hydroxy-ornithine (Fo-OH-Orn)	176.0797	158.0691	0	0	107
3	N-hydroxy-ornithine (OH-Orn)	148.0848	130.0742, 148.0848	2	0	93, 159
4	N-hydroxy-cyclo-ornithine (OH-cOrn)	130.0742	130.0742	2	0	93
5	N-hydroxy-lysine (OH-Lys)	162.1004	144.0898, 162.1004	0	not in NORINE	114, 232
6	5-Amino-N-hydroxypentan-1-amine (5AHA)	118.1106	118.1106	0	Not in NORINE	93
7	4-Amino-N-hydroxybutan-1-amine (4AHA)	104.095	104.095	0	Not in NORINE	48
8	3-Amino-N-hydroxypropan-1-amine (3AHA)	90.0793	90.0793	0	Not in NORINE	69
9	Citric acid (Cit)	192.0270	174.0164	0	Not in NORINE	44
10	Hydroxy-aspartic acid (OH-Asp)	149.0324	131.0218	0		45
11	2,3-Dihydroxybenzoic acid (di-OH-Bz)	154.0266	136.0160	0		39
12	Pyoverdine chromophore (ChrP)	277.1063	259.0957	0		52

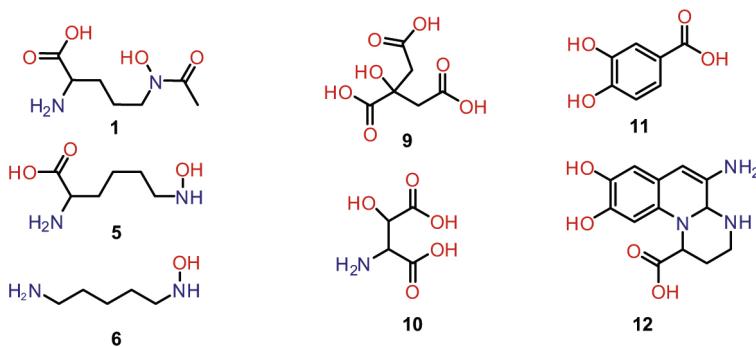
(\*) NORINE database with >1,100 nonribosomal peptides (NRPs): <http://bioinfo.lifl.fr/norine/>.

(\*\*) putative NRPS products.

(\*\*\*) METLIN MS/MS database with >70,000 high-resolution MS/MS spectra: <https://metlin.scripps.edu>.

**Table 1** Common iron-binding peptide monomers in known siderophores. Corresponding neutral-loss and fragment-ion searches can be used to mine LC-MS/MS datasets for siderophores. The mass for fragment ion searches can be obtained by adding a proton to the given neutral loss masses (+1.00783). Database searches in NORINE and METLIN were performed using a mass tolerance of  $\pm 0.005$  Da. Corresponding compound structures are shown in Figure 7.

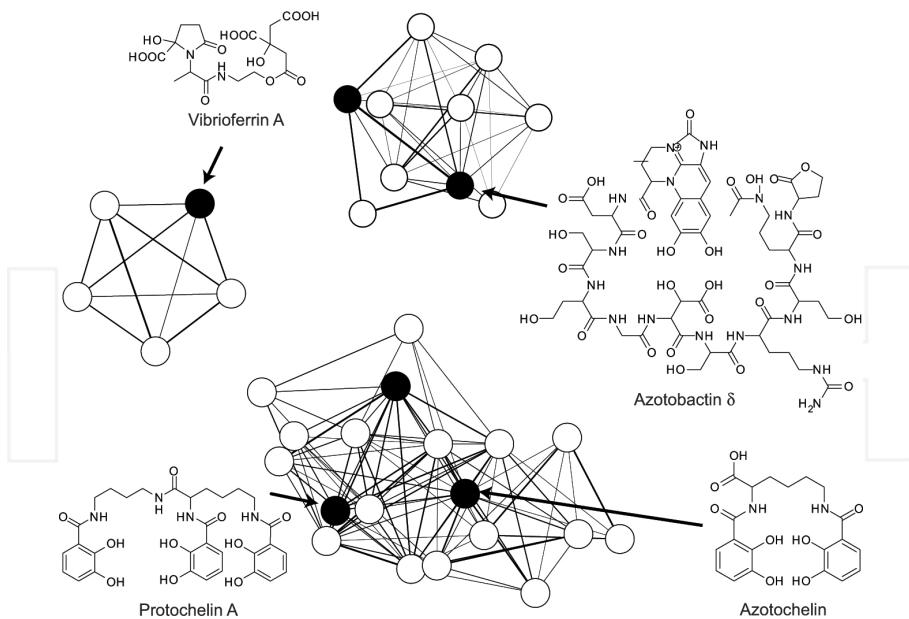
Application of neutral-loss and fragment-ion searches with the standards DFOB (containing 5AHA, compound 6), enterobactin (containing di-OH-Bz, compound 11), and amphibactin (containing Ac-OH-Orn, compound 1) readily reveal the parent ion  $m/z$  values of the standards. Application with a supernatant sample from the bacterium *A. vinelandii* yield a large number of siderophores related to its known catechol siderophores aminochelin, azotochelin, and protochelin as well as vibrioferrin, in agreement with previously reported results that were based on Fe stable isotope pattern screening of LC-MS data [27]. However, the *A. vinelandii* extract also contains a number of neutral-losses corresponding to siderophore substructures in Table 1, which upon further structural characterization are identified as false positives, originating from noise in the spectra. This further demonstrates that neutral-loss or fragment-ion searches alone are not sufficient for identification of siderophores. Nevertheless, they can provide a shortlist of candidate siderophore  $m/z$  values for structural characterization. If siderophore structures in a sample are known or expected (e.g., from genomic analyses), structure-specific fragments or neutral-losses can also be used to find related analogues. For example, the amphiphilic amphibactin siderophores yield fragmentation spectra with a common headgroup fragment:  $m/z = 450.219$  [33]. This  $m/z$  may be used to screen suspect samples for the presence of amphibactin-related siderophore species.



**Figure 7.** Common iron-binding substructures in siderophores, including hydroxamic acids (**1**, **5**, **6**),  $\alpha$ -hydroxycarboxylic acids (**9**, **10**), and catechols and related structures (**11**, **12**).

### 6.3. Siderophore MS/MS similarity networks

Bacteria are known to often produce suites of structurally closely related siderophores [17]. MS/MS similarity networks can give an overview of which precursors in a list of candidate siderophores are structurally independent, potentially originating from separate siderophore gene clusters, and which are likely to be structural analogues. An MS/MS similarity network for *A. vinelandii* distinguishes the three main independent siderophore structures that this bacterium produces: dihydroxybenzoic acid containing siderophores, vibrioferrin-type citrate containing siderophores, and azotobactin related siderophores as reported previously [27] (Figure 8). The MS/MS network facilitates efficient structural assignments and structural refinement (see below).



**Figure 8.** Siderophore MS/MS molecular network from the supernatant of *A. vinelandii* cultures, modified from [27]. Each node represents an individual siderophore and a corresponding consensus MS/MS spectrum; edge thicknesses represent the cosine similarity. An edge weighted-force directed layout in Cytoscape was used to cluster closely related nodes.

#### 6.4. Application of MS/MS auto-convolution to siderophore analysis

The success of de novo structural analysis of nonribosomal peptides (NRPs) using spectral auto-convolution depends on the identification of most or all amino acids involved in the structure, thus requiring a database that contains all involved peptide monomers as well as a good coverage of fragments in the MS/MS [19, 34]. To use auto-convolution for siderophore analysis, a database with >300 siderophore structures was compiled along with a database of the peptide-monomers occurring in these structures (160 different structures with 51 different Fe binding monomers), most of which are not present in the NORINE database of nonribosomal peptides (<http://bioinfo.lifl.fr/norine/>). Auto-convolution spectra were previously applied to cyclic NRPs (see Section 5.3), in which an MS/MS experiment leads to ring opening, and MS3 creates additional fragmentation [19]. Because ring opening may occur at any peptide bond, the theoretical MS3 spectra are a superposition of spectra derived from all possible linear peptides (circular permutations).

A modified auto-convolution approach is used here for analysis of peptidic siderophores and applied to siderophore structures that can also be linear, branched, or partly cyclic. Before auto-convolution, the algorithm adds the parent ion as a peak to the MS/MS spectrum and its intensity is set equal to the most intense peak in the spectrum. This ensures that the all precursor neutral loss  $m/z$  values occur in the auto-convolution spectrum, even if the precursor

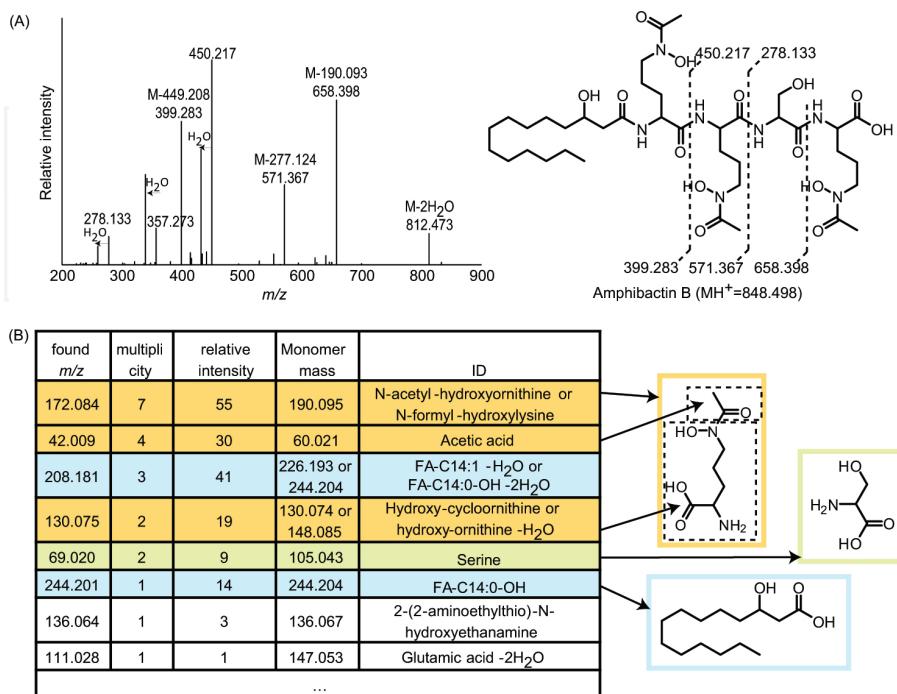
ion is not present as a peak in the MS/MS spectrum. In addition to the auto-convolution spectrum, the algorithm calculates the sum of the relative intensities of all fragment-ions associated with each  $m/z$  in the auto-convolution spectrum. The  $m/z$  values in the auto-convolution spectrum are then matched to masses in the siderophore database, taking into account possible neutral-losses of one or two H<sub>2</sub>O or NH<sub>3</sub>. Finally, neutral-charge masses of fragment-ions in the original MS/MS spectrum are calculated and also matched to the database of siderophore peptide monomers.

The relevance of database hits is judged by the multiplicity in the auto-convolution spectrum and the corresponding relative intensities of the MS/MS peaks involved. Illustrative results from the analysis of the siderophore amphibactin B are shown in **Figure 9**. The auto-convolution peaks with the highest multiplicity and highest relative intensities reveal all structural features of the molecule: the iron binding N-acetyl-hydroxyornithines, a serine, and the fatty acid tail. Combinations of monomers are then used as a fingerprint to search for possible related structures in the siderophore database. Four families of siderophores in the database contain the three possible substructures: amphibactins, aquachelins, marinobactins, and loihichelins. With this information, the amphibactin can be readily identified by reconstruction of the original MS/MS spectrum. A number of  $m/z$  differences shown in **Figure 9** have database matches with low multiplicity and relative intensity and do not relate to peptide monomers in the structure. One cause of false identifications of monomers can be spectral noise and contaminant peaks. To eliminate noise in the auto-convolution spectra, the analysis can be repeated with other related compounds in the siderophore MS/MS network and only those monomers prominent in a majority of spectra may be considered for structure proposals.

The approach was also successfully applied to the other siderophore standards used in this study: DFOB showed the iron-binding monomer 5AHA (Table 1) together with the succinic acid linker among the three substructures with the highest intensity and multiplicity. One potential peptide monomer ( $N^1,N^1$ -dimethyl- $N^5$ -acetyl- $N^5$ -hydroxy-ornithine,  $m = 200.1161$ ) was a false match as it has the same sum formula as the sum of 5AHA ( $m = 118.1106$ ) and succinic acid ( $m = 82.0055$  for succinic acid-2H<sub>2</sub>O). Searching the siderophore database for these monomers revealed ferrioxamines as most likely related structure. The cyclic enterobactin showed a prominent fragment with a mass corresponding to the iron binding dihydroxybenzoic acid groups in the structure as well as high multiplicity and intensity for the serines in the structure. The *A. vinelandii* supernatant contains three groups of structurally unrelated siderophores (**Figure 8**). Vibrio ferrins were associated with a strong auto-convolution peak for the iron-binding citric acid monomer in the molecule among a number of unrelated peptide monomers that were also matched. Protochelin, azotochelin, and related structures produced by *A. vinelandii* showed the characteristic dihydroxybenzoic acid together with the lysine or putrescine linkers contained in the structures. In contrast, azotobactin-related compounds in the supernatant did not show a clear siderophore signature, due to poor peptide fragmentation around the iron binding groups in the molecule.

When analyzing an unknown siderophore, the described auto-convolution approach can give confidence in the siderophore assignment: siderophores often contain three characteristic iron chelating peptide monomers for hexadentate iron coordination, which cause high multiplicity in the auto-convolution spectrum.

ties and relative intensities. A combination of peptide monomers in the structure can be used as a fingerprint to search the siderophore database for possible related structures, giving insight into the possible ‘siderophore class’ and aiding in the reconstruction of the original MS/MS spectrum to make a structure suggestion.



**Figure 9.** (A) MS/MS spectrum of the siderophore amphibactin B. (B) Results after application of a modified auto-convolution approach. Only auto-convolution  $m/z$  values are shown in the table that correspond to masses in a siderophore peptide monomer database. The  $m/z$  values with highest multiplicity and relative intensity reveal the structural features of this siderophore: the iron-binding N-acetyl-hydroxyornithines (orange), a serine (green), and a fatty acid tail (blue). A number of  $m/z$  values have database matches with low multiplicity and relative intensity and do not relate to peptide monomers in the structure (white).

## 7. Conclusions

The basic tools and considerations introduced in this chapter provide insight into the systematic analysis of LC-MS/MS fragmentation data for the structural characterization of small molecules and demonstrate how this can be performed within MATLAB. Since many researchers are familiar with MATLAB, this environment provides a low-barrier entry point and facilitates the creation of new strategies and tools to exploit the full power of modern high-resolution LC-MS/MS for structural interrogation. MATLAB facilitates data handling, manip-

ulation, and the implementation of graphical user interfaces to serve as a platform for the visualization, pre-processing, and analysis of LC-MS/MS data.

The discussed methods were applied in a new workflow for the discovery and structural characterization of siderophores by high-resolution LC-MS/MS. Using a database with siderophore structures, characteristic neutral-loss and fragment-ion masses were identified to mine LC-MS/MS data for potential siderophores. MS/MS siderophore networks in combination with a modified MS/MS auto-convolution approach revealed siderophore peptide monomers and corresponding siderophore families. This information was key to structure assignments by reconstruction of the original MS/MS spectrum. The tools and approaches outlined here may also be adapted to explorations of other classes of complex small molecules.

## Acknowledgements

We thank Yubo Li for helping with the assembly of the siderophore database and Alison Butler (UC Santa Barbara) for providing the amphibactin samples. Financial support for O.B. was provided by the Grand Challenge Program of the Princeton Environmental Institute.

## Supporting information

The MATLAB siderophore analysis software ('MS2Browser') and the siderophore database are available for download on SourceForge (<https://sourceforge.net/projects/ms2browser/>).

## Author details

Oliver Baars<sup>1\*</sup> and David H. Perlman<sup>2</sup>

\*Address all correspondence to: [obaars@princeton.edu](mailto:obaars@princeton.edu)

1 Department of Geosciences, Princeton University, Princeton, NJ, USA

2 Department of Chemistry, Princeton University, Princeton, NJ, USA

## References

- [1] Patti GJ, Yanes O, Siuzdak G. Metabolomics: the apogee of the omics trilogy. *Nature Reviews Molecular Cell Biology*. 2012;13(4):263–9.
- [2] Mastrangelo A, Armitage EG, Garcia A, Barbas C. Metabolomics as a tool for drug discovery and personalised medicine. A review. *Current Topics in Medicinal Chemistry*. 2014;14(23):2627–36.

- [3] Beger R. A review of applications of metabolomics in cancer. *Metabolites*. 2013;3(3):552.
- [4] Zwiener C, Frimmel FH. LC-MS analysis in the aquatic environment and in water treatment technology—a critical review. Part II: Applications for emerging contaminants and related pollutants, microorganisms and humic acids. *Analytical and Bioanalytical Chemistry*. 2004;378(4):862–74.
- [5] Cheng X, Hochlowski J. Current application of mass spectrometry to combinatorial chemistry. *Analytical Chemistry*. 2002;74(12):2679–90.
- [6] Castillo S, Gopalacharyulu P, Yetukuri L, Orešić M. Algorithms and tools for the preprocessing of LC-MS metabolomics data. *Chemometrics and Intelligent Laboratory Systems*. 2011;108(1):23–32.
- [7] Sugimoto M, Kawakami M, Robert M, Soga T, Tomita M. Bioinformatics tools for mass spectroscopy-based metabolomic data processing and analysis. *Current Bioinformatics*. 2012;7(1):96–108.
- [8] Wolfender JL, Marti G, Thomas A, Bertrand S. Current approaches and challenges for the metabolite profiling of complex natural extracts. *Journal of Chromatography A*. 2015;1382:136–64.
- [9] Pluskal T, Castillo S, Villar-Briones A, Oresic M. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics*. 2010;11:395.
- [10] Benton HP, Wong DM, Trauger SA, Siuzdak G. XCMS2: processing tandem mass spectrometry data for metabolite identification and structural characterization. *Analytical Chemistry*. 2008;80(16):6382–9.
- [11] Scheubert K, Hufsky F, Böcker S. Computational mass spectrometry for small molecules. *Journal of Cheminformatics*. 2013;5:12.
- [12] Xiao JF, Zhou B, Ressom HW. Metabolite identification and quantitation in LC-MS/MS-based metabolomics. *Trends in Analytical Chemistry*. 2012;32:1–14.
- [13] Kind T, Fiehn O. Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinformatics*. 2006;7(1):1–10.
- [14] Stein SE. Mass spectral reference libraries: an ever-expanding resource for chemical identification. *Analytical Chemistry*. 2012;84(17):7274–82.
- [15] Watrous J, Roach P, Alexandrov T, Heath BS, Yang JY, Kersten RD, et al. Mass spectral molecular networking of living microbial colonies. *Proceedings of the National Academy of Sciences*. 2012;109(26):E1743–E52.
- [16] Kind T, Fiehn O. Advances in structure elucidation of small molecules using mass spectrometry. *Bioanalytical Reviews*. 2010;2(1–4):23–60.

- [17] Hider RC, Kong XL. Chemistry and biology of siderophores. *Natural Product Reports.* 2010;27(5):637–57.
- [18] Holman JD, Tabb DL, Mallick P. Employing ProteoWizard to convert raw mass spectrometry data. *Current Protocols in Bioinformatics.* 2014;46:13.24.1–9.
- [19] Ng J, Bandeira N, Liu W-T, Ghassemian M, Simmons TL, Gerwick WH, et al. DerePLICATION and de novo sequencing of nonribosomal peptides. *Nature Methods.* 2009;6(8):596–9.
- [20] Frank AM, Bandeira N, Shen Z, Tanner S, Briggs SP, Smith RD, et al. Clustering millions of tandem mass spectra. *Journal of Proteome Research.* 2008;7(1):113–22.
- [21] Yang X, Neta P, Stein SE. Quality control for building libraries from electrospray ionization tandem mass spectra. *Analytical Chemistry.* 2014;86(13):6393–400.
- [22] Tsugawa H, Cajka T, Kind T, Ma Y, Higgins B, Ikeda K, et al. MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nature Methods.* 2015;12(6):523–6.
- [23] Stein SE. An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. *Journal of the American Society for Mass Spectrometry.* 1999;10(8):770–81.
- [24] Brown NL, Stoyanov JV, Kidd SP, Hobman JL. The MerR family of transcriptional regulators. *FEMS Microbiology Reviews.* 2003;27(2–3):145–63.
- [25] Stein SE. Chemical substructure identification by mass spectral library searching. *Journal of the American Society for Mass Spectrometry.* 1995;6(8):644–55.
- [26] Pevzner PA, Dančík V, Tang CL. Mutation-tolerant protein identification by mass spectrometry. *Journal of Computational Biology.* 2000;7(6):777–87.
- [27] Baars O, Zhang X, Morel FM, Seyedsayamdst MR. The siderophore metabolome of *Azotobacter vinelandii*. *Applied and Environmental Microbiology.* 2015;82(1):27–39.
- [28] Brodbelt JS. Ion activation methods for peptides and proteins. *Analytical Chemistry.* 2016;88(1):30–51.
- [29] Chaturvedi KS, Henderson JP. Pathogenic adaptations to host-derived antibacterial copper. *Frontiers in Cellular and Infection Microbiology.* 2014;4:3.
- [30] Roepstorff P, Fohlman J. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomedical Mass Spectrometry.* 1984;11(11):601.
- [31] Caboche S, Pupin M, Leclère V, Fontaine A, Jacques P, Kucherov G. NORINE: a database of nonribosomal peptides. *Nucleic Acids Research.* 2008;36:D326–D31.
- [32] Baars O, Morel FMM, Perlman DH. ChelomEx: isotope-assisted discovery of metal chelates in complex media using high-resolution LC-MS. *Analytical Chemistry.* 2014;86(22):11298–305.

- [33] Martinez JS, Carter-Franklin JN, Mann EL, Martin JD, Haygood MG, Butler A. Structure and membrane affinity of a suite of amphiphilic siderophores produced by a marine bacterium. *Proceedings of the National Academy of Sciences*. 2003;100(7):3754–9.
- [34] Mohimani H, Liu W-T, Yang Y-L, Gaudêncio SP, Fenical W, Dorrestein PC, et al. Multiplex de novo sequencing of peptide antibiotics. *Journal of Computational Biology*. 2011;18(11):1371–81.

INTECH

INTECH