Computational and Statistical Methods for Mass Spectrometry Data Analysis

summary of the personal accomplishments

Mateusz Krzysztof Łącki

Department of Mathematics, Informatics, and Mechanics, University of Warsaw

Mass Spectrometry is a subfield of the Analytical Chemistry that studies and develops instruments useful for analyzing the molecular content of samples. The instruments, that are called mass spectrometers, have been developed by Joseph J. Thomson just before the First World War. The output of a mass spectrometer – a mass spectrum – is a histogram: one the x-axis it reports the observed mass-to-charge ratios. At each reported mass-to-charge ratios there is a peak, whose height, referred to as intensity, is a function of the number of molecules observed by the detector with that given mass-to-charge ratio. Below, we show how a mass spectrum might look like.

Consider a compound with a fixed formula, such as the human insuline, $\text{C}_{520}\text{H}_{817}\text{N}_{139}\text{O}_{147}\text{S}_{8}$. What kind of signal can it produce in a mass spectrum? To answer this question, we have to consider several factors. First of all, all atoms in the above formula can assume one of the multiple isotopic variants, from a set that is different for each element. Table 1 summarizes these possible variants. The reported abundances reported there are the marginal probabilities of the isotopic
Table 1: Masses and Frequencies of isotopes of elements that make up proteins.

Source: Brand et al. 1.

<table>
<thead>
<tr>
<th>Isotope</th>
<th>Mass</th>
<th>Abundance</th>
</tr>
</thead>
<tbody>
<tr>
<td>$^1$H</td>
<td>1.0078</td>
<td>0.9999</td>
</tr>
<tr>
<td>$^2$H (D)</td>
<td>2.0141</td>
<td>0.0001</td>
</tr>
<tr>
<td>$^{12}$C</td>
<td>12.0000</td>
<td>0.9892</td>
</tr>
<tr>
<td>$^{13}$C</td>
<td>13.0034</td>
<td>0.0108</td>
</tr>
<tr>
<td>$^{14}$N</td>
<td>14.0031</td>
<td>0.9964</td>
</tr>
<tr>
<td>$^{15}$N</td>
<td>15.0001</td>
<td>0.0036</td>
</tr>
<tr>
<td>$^{16}$O</td>
<td>15.9949</td>
<td>0.9976</td>
</tr>
<tr>
<td>$^{17}$O</td>
<td>16.9991</td>
<td>0.0004</td>
</tr>
<tr>
<td>$^{18}$O</td>
<td>17.9992</td>
<td>0.0021</td>
</tr>
<tr>
<td>$^{32}$S</td>
<td>31.9721</td>
<td>0.9499</td>
</tr>
<tr>
<td>$^{33}$S</td>
<td>32.9715</td>
<td>0.0075</td>
</tr>
<tr>
<td>$^{34}$S</td>
<td>33.9679</td>
<td>0.0425</td>
</tr>
<tr>
<td>$^{36}$S</td>
<td>35.9671</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

variants: they describe the possible isotopic variant of only one atom. By far the easiest way to construct a joint probability measure out of these marginals is to assume that isotopic variants of different atoms are mutually independent. In certain situations, such as the isotopic labeling, one would certainly have to modify that model, as these could introduce some non-trivial dependence between the isotopic variants.

Mass spectrometer does not distinguish compounds with the same number of isotopic variants. For instance, if one of the 817 hydrogen atoms of human insulin is deuterium, then, based on mass spectrum alone, there is no way of telling which particular atom was the heavier one. The observed signal depends only upon counts of different isotopic variants. Similarly to how chemical formulas such as C$_c$H$_h$N$_n$O$_o$S$_s$ abstract from spatial composition, we may introduce a more detailed description of the isotopic content of a molecule, such as

\[ \text{iso} = 12C_{c_0}^{13}C_{c_1}^*^{1}H_{h_0}^*^{2}H_{h_1}^{14}N_{n_0}^{15}N_{n_1}^{16}O_{o_0}^{17}O_{o_1}^{18}O_{o_2}^{32}S_{s_0}^{33}S_{s_1}^{34}S_{s_2}^{36}S_{s_3}. \]

Above, $c_0$ stands for the number of $^{12}$C isotopes within the molecule, $c_1$ – number of $^{13}$C isotopes, and so on. This is essentially what we call an *isotopologue*. This definition coincides with that provided by the International Union of Pure and Applied Chemistry ². The independence assump-
tions suggest that the probability of observing an isotopologue is that of a product of multinomial distributions, each for one element, or

\[
p_{\text{iso}} = \binom{c}{c_0, c_1} p\left(1^{12}\text{C}\right)^{c_0} p\left(1^{13}\text{C}\right)^{c_1} \binom{h}{h_0, h_1} p\left(1^{1}\text{H}\right)^{h_0} p\left(1^{2}\text{H}\right)^{h_1} \binom{n}{n_0, n_1} p\left(1^{14}\text{N}\right)^{n_0} p\left(1^{15}\text{N}\right)^{n_1} \times \binom{o}{o_0, o_1, o_2} p\left(1^{16}\text{O}\right)^{o_0} p\left(1^{17}\text{O}\right)^{o_1} p\left(1^{18}\text{O}\right)^{o_2} \binom{s}{s_0, s_1, s_2, s_3} p\left(1^{32}\text{S}\right)^{s_0} p\left(1^{33}\text{S}\right)^{s_1} p\left(1^{34}\text{S}\right)^{s_2} p\left(1^{36}\text{S}\right)^{s_3}.
\]

The mass of \(\text{iso}\) is given by multiplying counts of different isotopes times their masses (Table 1), \(m_{\text{iso}} = m\left(1^{12}\text{C}\right)c_0 + m\left(1^{13}\text{C}\right)c_1 + \cdots + m\left(1^{36}\text{S}\right)s_3\). The set of all pairs \((p_{\text{iso}}, m_{\text{iso}})\) that corresponds to one chemical formula \(\text{C}_c \text{H}_h \text{N}_n \text{O}_o \text{S}_s\) makes up the isotopic fine structure. The isotopic fine structure is typically used directly to model the signal in the instrument.

**Isotopic Calculations**

In Chapter 1 of my dissertation, I describe an efficient and elegant way to quickly generate high coverage subsets of the isotopic fine structure. What is understood by a high coverage set, is a set of isotopologues with joint probability not smaller than some threshold \(P\). We call that set an optimal \(P\)-set. The algorithm that retrieves the elements of the optimal \(P\)-set relies on two crucial properties of the multinomial distribution: its concentration of measure\(^3\) and unimodality. Measure concentration assures that a relatively small number of configurations carry most of that distribution’s probability mass. Unimodality, defined in this context in terms of the connectedness of the set of local maxima, assures that it is easy to visit all configurations in a descending order. These features also apply to the product of the multinomial distributions, which is precisely the probability measure that appears in the context of the isotopic calculations.

What is more, an algorithm that uses a simple FIFO queue is shown to efficiently perform the calculations. In particular, we show that the generation of consecutive peaks in an optimal \(P\)-set requires a linear number of calculations with respect to its size, which is optimal. To establish the time complexity, we make use of the Central Limit Theorem that we apply to the multinomial distribution, see Figure 1. We show, that the number of isotopologues inside the optimal \(P\)-set
can be approximated well by formula

\[
M = \frac{q\chi^2(k)(p)^\frac{k}{2}}{C} \frac{\pi^{k/2}}{\Gamma(k/2 + 1)} \prod_{e \in \mathcal{E}} \left( \prod_{e \in \mathcal{E}} \sum_{j=0}^{i_e-1} \right) \prod_{e \in \mathcal{E}} \left( \prod_{j=0}^{i_e-1} \bar{d}_{e,j} \right).
\]

In particular, from the above fact one can deduce that by applying the joint probability thresholding we obtain a set of configuration that is of order of a square root of the total number of configurations that make up the isotopic distribution. In most cases, this is why the calculations can be carried out, even for compounds as big as titin, \( \text{C}_{160719}\text{H}_{270464}\text{N}_{45688}\text{O}_{52237}\text{S}_{911} \). What is more, the implementation of the algorithm significantly outperforms other existing isotopic calculators.

The applications of the ideas that resulted in this algorithm go beyond mass spectrometry, and their use is now investigated in statistics and stochastic simulation.

**Deconvolution of Mass Spectra and Ion Statistics**

The results of the IsoSpec only describe the isotopic variants attainable by only one molecule. Nonetheless, if one assumes that different ions appear independently in the instrument, then that distribution is directly proportional to the expected shape of the isotopic distribution. Therefore, it can be directly applied to the problem of establishing the presence of a signal coming from one

---

\(^1\)Of course, we bring up this example only to illustrate the algorithms capabilities. It would not make too much sense to introduce to the instrument a compound that big.
source of ions, via non-negative regression schemes similar to that proposed by Slawski et al.\textsuperscript{4}. We further generalize that idea in a project called MassTodon, described in the next section.

In reality, the intensity observed in a mass spectrometer is a function of a high number of ions. It has been long since established that the distribution of the number of ions should follow a Poisson distribution\textsuperscript{5,6}. The theoretical argument behind this is as simple: the number of ions that actually make it to the detector is very limited. Assume that ions move independently throughout the instrument and that each one has some chance of reaching the detector. Then, the number of ions that made it to the detector can be directly modeled by a binomial model. That distribution is well approximated by the Poisson distribution if the probability of the success, i.e. reaching the detector is very small. That is why the Poisson distribution is also sometimes referred to as the distribution of rare events. Chapter 4 describes our attempts to merge the concepts of the isotopic distribution with the Poissonian ion statistics into one tool that we call \texttt{MassOn}.

\texttt{MassOn} tackles two important problems in mass spectrometry signal processing:

1. the deconvolution of a compound signal

2. the estimation of the number of observed ions

The nature of that first problem lies in the limited capability of a mass spectrometer to resolve close mass-to-charge ratios of different molecular species. In particular, it can very well happen that isotopologues of different molecular species can contribute to the height of the same peak. This idea is sketched in Figure 2.

The second problem stems from the fact that many instruments do not directly detect the number of ions, but rather some function thereof. The problem of estimating the shape of that function is of great relevance, as the number of observed ions appears in most of the expressions involving the standard deviations of statistics derived from the theoretical mass spectrum. In particular, if one assumes that the recorded peak height indeed results from an independent motion of ions close to the detector, then the standard deviation of that peak is a function of the square root of the overall number of ions. Both Chapters 2 and 3 describe other important statistics that
rely on the specification of the recorded number of ions. MassOn tackles both these problems in a fully Bayesian setting relying on a \textit{data augmented} Gibbs sampling scheme.

In that model, we treat molecular species as conditionally independent Poisson processes that populate the mass over charge half-line. Every process can be assigned an intensity measure $\mu_m$, that relates the isotopic distribution to the mass-over-charge values,

$$
\mu_m(A|\Lambda_m) = \Lambda_m \sum_{mz \in MZ_m} p_{mz} g(A - mz). \tag{1}
$$

Above, $\Lambda_m$ is the total intensity of the species $m$, $p_{mz}$ is the probability of observing an ion with mass-to-charge ratio $mz$. All possible mass-to-charge ratios for molecular species $m$ are gathered in set $MZ_m$. The above processes are conditionally independent given the values of intensities $\Lambda_m$ and the marginal frequencies of isotopes, that are used to calculate $p_{mz}$.

These intensity measures describe, on average, the ionic content of the mass spectrum at different mass ranges. These numbers are not directly observed in the mass spectrum. We assume that the observed peak intensity $I_A$ in a mass-to-charge range $A = [a_{\text{min}}, a_{\text{max}}]$ is a polynomial function of the real number of ions. The model, apart from deconvoluting the isotopic distributions, can also find the \textit{a posteriori} distribution of the \textit{ion-intensity} coefficients of that polynomial.
Understanding Reaction Pathways

Another limitation of any mass spectrometer is that substances with the same chemical formula and different structures cannot be told apart. In particular, this is the case of two post-translationally modified proteins that have the same modification that could be found on more than one residue. The spatial positioning of a modification is critical for the folding of the protein, and thus – its function. To position a PTM, one has to use more specific techniques, fragmentation being one of them. Ions can be fragmented either outside the instrument, via proteolytic digestion, or inside the instrument. Two prominent ways of inducing fragmentation inside the instrument are the Collisional Induced Dissociation (CID) and the Electron Transfer Dissociation (ETD). The first one consists in heating up the sample cations by exposing them to collisions with some inert gas. This method produces more noisy spectra, as different parts of the molecule detach due to their increased internal motion. ETD is much more subtle technique: it consists of an ion-ion reaction between the sample cations and anions, each carrying a radical – an electron in a higher energy state. The meeting between these ions is deemed to result in four possible outcomes:

- the transfer of electron from the anion to the cation resulting in the dissociation of the cation – the proper ETD
- the transfer of electron that does not result in any dissociation – ETnoD
- the ETD dissociation followed by a subsequent hydrogen transport – HTR
- the proton transfer reaction – PTR

Table 2 shows the details of the considered reactions.

To study the products of these fragmentation, we devised an approach named MassTodon. Chapter 2 provides a detailed explanation of the approach we take to study electron transfer driven reactions. The presented workflow can find the products of these reactions in the spectrum. In order to do this, MassTodon:
1. generates the formulas of all possible reaction products
2. investigates their presence in mass spectrum
3. deconvolutes different isotopic distributions using constrained quadratic programming
4. pairs the intensities of the matching c and z ions in order to estimate the fragmentation probabilities and the branching ratio

While performing point 2 in the above listing, MassTodon applies IsoSpec to generate the expected shapes of the signal. These are also used for deconvolution. To efficiently perform all the calculations, MassTodon relies on the concept of the *Deconvolution Graph*. The above graph representation offers the possibility to naturally parallelize the deconvolution calculations, that can be performed independently on different connected components. Figure 3 shows how to relate the *deconvolution graph* to the observed mass spectrum.

MassTodon outputs the estimates of joint intensities of each chemical formula it found from the set of potential reaction products and substrates. It can also estimate the probabilities with which different reactions occurred in one experiment. This simplifies the comparison of different mass spectra, offering a possibility to better study the influence of different instrumental settings upon the sample; finally, it also simplifies the comparison of different instruments. In particular, MassTodon has been already used to estimate the branching ratio (the odds ratio between ETnoD and PTR) that has been found to correlate well with the collision cross-section of proteins.

Table 2: Considered chemical reactions. M stands for either a precursor ion or a fragment ion. The HTR reaction can happen only after ETD and consists in the transfer of a hydrogen atom from the c to the z fragment.

<table>
<thead>
<tr>
<th>Reaction</th>
<th>Product 1</th>
<th>Product 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>PTR</td>
<td>[M + nH](^{n+}) + A(^{+})</td>
<td>([M + (n-1)H](^{(n-1)+}) + AH</td>
</tr>
<tr>
<td>ETnoD</td>
<td>[M + nH](^{n+}) + A(^{+})</td>
<td>([M + nH](^{(n-1)+}) + A</td>
</tr>
<tr>
<td>ETD</td>
<td>[M + nH](^{n+}) + A(^{+})</td>
<td>([c + xH](^{x+}) + [z + (n - x)H](^{(n-x-1)+}) + A</td>
</tr>
<tr>
<td>HTR</td>
<td>[c + xH](^{x+})</td>
<td>([c + (x - 1)H](^{x+})</td>
</tr>
<tr>
<td></td>
<td>[z + (n - x)H](^{(n-x-1)+})</td>
<td>([z + (n - x + 1)H](^{(n-x-1)+})</td>
</tr>
</tbody>
</table>

8
Unveiling Reaction Kinetics

With the estimates of the intensity of particular molecular species at hand, one can pose more specific questions on the nature of the chemical process that could generate them. A natural approach in this context is to apply the mathematical apparatus provided by the theory of reaction kinetics. In particular, we adapted an approach based on a dynamic stochastic Petri net proposed by Gambin and Kluge, schematically shown in Figure 4.

The model offers a parametrization of the process in terms of a set of reaction rates, specific for different reactions. Due to the nature of the process, the considered net does not have any self-loops, which significantly simplifies the dynamics of the process. In particular, one can derive recursive formulas for the average amounts of ions at each possible place in the net, i.e. of each molecular species that can be observed during the reactions. These formulas depend on the considered parameters. By applying a gradient-free optimization algorithm, we try to find parameters that minimize the distance from the estimates of the total intensities of different molecular species that appear in the Petri net.

The developed tool, called ETDetective, is integrated with MassTodon. The model provides a much more comprehensive way to study the outcomes of the electron-driven reactions, providing
Figure 4: A model of the ETD reaction. (a) A fragment of the reaction graph for a triply charged precursor. The molecular species are depicted in pale grey and the reactions in dark grey. The skull represents the cemetery. The reaction graph serves as a board for tokens that represent the numbers of molecules of a given species, depicted as black circles. Only one ETD transition has been shown for clarity of the image. (b) During each reaction, a token disappears on the substrate side and product tokens appear: one in the case of ETnoD and PTR, two in the case of ETD).

direct access to reaction rates that are easily interpretable by the chemists. Chapter 3 describes in detail the approach we follow to perform these calculations.

Publications in the field of Mass Spectrometry


References


