4 Database Searching and Aligning Mass Spectra

> "If it looks like a duck, and quacks like a duck, we have at least to consider the possibility that we have a small aquatic bird of the family Anatidae on our hands." (Douglas Adams, Dirk Gently's Holistic Detective Agency)

GIVEN A MEASURED SAMPLE SPECTRUM and a database with reference spectra, there are two questions to be answered: Which reference spectrum matches best with the measured? And how certain are we, that our identification is correct? In this chapter, we will focus on the first question; the second question will be addressed in Chapters 5 and 6.

Our presentation in this chapter will again focus on identifying peptides using tandem mass spectrometry. It must be understood that many of the concepts introduced here, can also be used in quite different contexts such as metabolite identification (Chapter 13) or glycan *de novo* sequencing (Chapter 14). Focusing on peptides, will help us to fill our theoretical concepts with some "meat". Also note that reference spectra can be computed on the fly for peptide identification, using a protein sequence database.

In the following, we assume that $\mathscr{M}$ is the *reference spectrum* we want to compare against; and that $\mathscr{M}'$ is the *measured spectrum* of our sample using an MS instrument. For the ease of presentation, we assume that both $\mathscr{M}$ and $\mathscr{M}'$ are sets of masses. In fact, we can easily add more "peak attributes" to this framework without having to change the formal presentation: We can think of these attributes as maps from the set of masses, to some set representing the possible attribute states. One such attribute that we will make use of repeatedly, are peak in intensities in the measured spectrum. For the reference spectrum, a possible peak attribute is the ion series the peak stems from.

[TODO: INTRODUCE PSM]

From the conceptual side, the problem of searching for something in a database is not as intellectually challenging as *de novo* sequencing, where we search through the much larger space of *all* possibilities. Algorithmically, there has been not much progress to attack, say, PTMs for peptide database searching. As a funny twist, it turns out that some approaches for accelerated peptide database searching, rely heavily on spectrum graphs and other ideas from the previous chapter; see Sec. 16.3. These approaches speed up tasks such as searching with PTMs, or searching for batches of spectra, and are currently the only ones conceptually convincing.

4.1 Matching mass spectra

Given a protein string, it is quite easy to simulate, say, tryptic digestion *in silico*, see Exercise 1.1. But it is similarly easy to simulate the tandem mass spectrum of a peptide — at least, if we assume some simple model of peptide fragmentation, such as the one from the previous chapter, or the one presented in Sec. 4.4 and 4.5. In fact, we have implicitly "simulated" such

peptide tandem mass spectra in the previous section. We leave the details to the reader, see Exercises 4.1 and 4.7.

In Chapter 2, we have implicitly introduced a simple approach to compare two mass spectra: We did so by counting the peaks that occur both in the measured spectrum $\mathcal{M}'$ and in the reference spectrum $\mathcal{M}$. This number will be called *peak counting score* in the following, but goes under many different names in the literature. The idea behind this, is that the measured spectrum is fixed, whereas we are searching for a best match in the database. As introduced in Sec. 2.5.6, we have to allow for some mass deviation $\varepsilon > 0$ between the masses of measured and reference peak. In the following, we will also look at other ways to compute a *score* for the reference spectrum $\mathcal{M}$ by comparing it to the fixed measured spectrum $\mathcal{M}'$.

What exactly do we mean with "counting common peaks"? In fact, there are at least three different interpretations:

1. We want to match pairs of peaks: That is, every peak in the reference spectrum $\mathcal{M}$ can be matched with at most one peak in the measured spectrum $\mathcal{M}'$, and vice versa, to contribute towards the score.

2. Each peak in the reference spectrum $\mathcal{M}$ can be matched with at most one peak in the measured spectrum $\mathcal{M}'$; but a peak in the measured spectrum $\mathcal{M}'$ may be matched to many peaks in the reference spectrum $\mathcal{M}$.

3. Each peak in the measured spectrum $\mathcal{M}'$ can be matched with at most one peak in the reference spectrum $\mathcal{M}$; but a peak in the reference spectrum $\mathcal{M}$ may be matched to many peaks in the measured spectrum $\mathcal{M}'$.

Intuitively, the first interpretation appears to be the most "natural"; but it turns out that the second interpretation is also quite reasonable in many applications. We will call the first a *one-to-one matching*, and the second a *many-to-one* matching. In contrast, the third interpretation should hardly ever be relevant in applications. We will discuss this later and, for the moment, concentrate on the one-to-one matching case.

Example 4.1. We now give an example meant to demonstrate various problems of the peak counting score. Assume that we have measured the spectrum

$$\mathcal{M}' = \{200, 300, 500, 515, 700\}$$

and we want to compare it against a set of reference mass spectra in our database, namely:

$$\begin{aligned}
\mathcal{M}_1 &= \{100, 175, 350, 480, 490, 550\} \\
\mathcal{M}_2 &= \{200, 270, 300, 500\} \\
\mathcal{M}_3 &= \{205, 505, 705, 850\} \\
\mathcal{M}_4 &= \{190, 310, 490, 710\} \\
\mathcal{M}_5 &= \{100, 150, 200, 250, \ldots, 600, 650, 700\}
\end{aligned}$$

Assume that $\varepsilon = 10$ is the mass deviation that we believe to be reasonable. Now, we find that $\mathcal{M}_1$ has one peak in common with $\mathcal{M}$; both $\mathcal{M}_2$ and $\mathcal{M}_3$ have three; and both $\mathcal{M}_4$ and $\mathcal{M}_5$ have four peaks in common.

What are the problems with the peak counting score in Example 4.1? First, changing parameter ε slightly can dramatically change the score. For example, spectrum $\mathscr{M}_4$ has score four for $\varepsilon = 10$, but if we instead choose $\varepsilon = 9.9$ then the peak counting score decreases to zero. But also for spectra that are not in this "critical zone", it is understood that $\mathscr{M}_2$ fits the observed data better than $\mathscr{M}_3$, but this is not reflected in the score. Finally, reference spectra with many peaks such as $\mathscr{M}_5$ get a better score, because the are more likely to hit a peak mass in $\mathscr{M}$ just by chance. We have observed this problem at the end of Sec. 2.5.3.

In view of this, it seems reasonable to score mass deviations a little more carefully. To this end, we assume that we are given some *mass scoring function* $f : \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \to \mathbb{R}$ that, for a pair of peaks at masses m (for the reference peak) and m' (for the measured peak), judges the similarity of these peaks based on their masses. For a mass scoring function to be true to the application, we may demand an additional property: If $M' < m' < m$ or $m < m' < M'$ holds, then $f(m,M') < f(m,m')$. Similarly, if $M < m < m'$ or $m' < m < M$ then $f(M,m') < f(m,m')$. A mass scoring function is called *strictly monotonical* if it satisfies these two conditions. A weaker condition is that $M' < m' < m$ or $m < m' < M'$ implies $f(m,M') \leq f(m,m')$, and that $M < m < m'$ or $m' < m < M$ implies $f(M,m') \leq f(m,m')$. In this case, the mass scoring function is called *monotonical*. For example, the peak counting score for any $\varepsilon > 0$ is monotonical but not strictly monotonical. The above are a quite reasonable conditions: For example, $f(M,m') > f(m,m')$ for $M < m < m'$ or $m' < m < M$ would imply that matching the measured peak at mass m' with the more distant reference peak M, is more sensible than matching it with the closer reference peak m.

Example 4.2. Let $g(m,m') := 1 - 2\left|m - m'\right|$ for $m,m' \in \mathbb{R}_{\geq 0}$. Then, g is a mass scoring function that is strictly monotonical. In particular, we have $g(m,m') \leq 1$ for all m,m'; $g(m,m') = 1$ if and only if $m = m'$; and $g(m,m') = 0$ for $\left|m - m'\right| = \frac{1}{2}$. **[TODO: BILD EINFUEGEN]**

We now assume that peak pairs are scored by some score function $\sigma : \mathscr{M} \times \mathscr{M}'$. Such a score function is usually derived from a mass scoring function, but can take into account other attributes such as intensities. An alignment of the mass spectra $\mathscr{M}$ and $\mathscr{M}'$ is a *matching* of the two sets, where a subset of $\mathscr{M}$ is bijectively mapped onto a subset of $\mathscr{M}'$. To penalize peaks that are not matched with any counterpart, we introduce a gap character ϵ. Here, $\sigma(\epsilon,m') \leq 0$ penalizes a missing peak $m' \in \mathscr{M}'$, whereas $\sigma(m,\epsilon) \leq 0$ penalizes an additional peak $m \in \mathscr{M}$. We define the score of a matching as:

$$\sum_{m \text{ matches } m'} \sigma(m,m') + \sum_{\text{missing peaks } m' \in \mathscr{M}'} \sigma(\epsilon,m') + \sum_{\text{additional peaks } m \in \mathscr{M}} \sigma(m,\epsilon) \tag{4.1}$$

Example 4.3. **[TODO: BEISPIEL EINFUEGEN]**

Crossing matchings (shown in Example 4.3) are not admissible, because they are physical nonsense. In fact, a strictly monotonical mass scoring function will never result in a crossing matching:

Lemma 4.1. *Given two mass spectra $\mathscr{M},\mathscr{M}'$ and a strictly monotonical mass scoring function $f : \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \to \mathbb{R}$. Assume that the optimal alignment of $\mathscr{M}$ and $\mathscr{M}'$ under the scoring $\sigma(m,m') := f(m,m')$ aligns the masses $\mathscr{A} \subseteq \mathscr{M} \times \mathscr{M}'$. Then, for two matched mass pairs (m_1,m_1') and (m_2,m_2') from $\mathscr{A}$, we have $m_1 < m_2$ if and only if $m_1' < m_2'$.*

See Exercise 4.8 for the proof of the lemma. The optimal matching can be found by aligning the spectra, so we use dynamic programming for the table $D[1\ldots n, 1\ldots n']$ with $n := |\mathscr{M}|$ and

$n' := |\mathcal{M}'|$. We initialize $D[0,0] = 0$, $D[i,0] = D[i-1,0] + \sigma(m_i, \epsilon)$ for $i = 1,\ldots,n$, and $D[0,j] = D[0,j-1] + \sigma(\epsilon, m'_j)$. We use the following recurrence to fill the table:

$$D[i,j] = max \begin{cases} D[i-1,j] + \sigma(m_i, \epsilon) \\ D[i-1,j-1] + \sigma(m_i, m'_j) \\ D[i,j-1] + \sigma(\epsilon, m'_j) \end{cases} \tag{4.2}$$

The score of an optimal alignment between Obviously, the method requires $O(n \cdot n')$ time and memory. After filling the matrix, the optimal score can be found in entry $D[n,n']$. To find the optimal alignment we use backtracking through D. Consider the measured spectrum $\mathcal{M}$ and reference spectrum **[TODO: WHICH?]** from Example 4.1: Using the mass scoring function from Example 4.2 with gap penalty -1, the best alignment has score 1. In application, the optimal alignment can usually be found much faster than the worst-case running times suggests: but is faster in application normally. For example if $\sigma(m,m') < \sigma(m,\epsilon) + \sigma(\epsilon,m')$ matching m and m' causes the optimal alignment in no case. This banded estimation needs only linear time and memory.

4.2 Fundamentals of scoring and mass accuracies

We will now prepare a "base stock" for scoring the agreement of two mass spectra and, in particular, scoring a measured spectrum against a reference spectrum. Our main ingredients will be "mass differences" and "peak intensities." Other ingredients and flavors can be added at your own choice.

For our score, we will use log odd scores, as defined in statistics: We want to differentiate between two statistical models, one for our hypothesis and one for the background. Here, we look at a pair of peaks, one from the measured spectrum and one from the reference spectrum, that have been matched by our spectrum alignment algorithms. Now, the two models are "the measured peak is an incorporation of the reference peak" vs. "the measured peak is simply noise, and has nothing to do with the reference peak."

Odd scores are used to differentiate between the two models, by computing the ratio

$$odd\,score = \frac{\mathbb{P}(D|H_1)}{\mathbb{P}(D|H_0)}$$

where D is the observed data (the peak in the measured spectrum), H_1 is our hypothesis (the measured peak belongs to the reference peak), and H_0 is the null model (the measured peak is noise). For $odd\,score > 1$ we would accept the model H_1, and for $odd\,score < 1$ the null model H_0 is more likely.

Log odd scores do pretty much the same as odd scores:

$$log\,odd\,score = \log \frac{\mathbb{P}(D|H_1)}{\mathbb{P}(D|H_0)} \tag{4.3}$$

Here, the logarithm can be computed to an arbitrary (but fixed) basis, such as the natural logarithm with basis e. For $\log_2$ the resulting log odd scores are called *bit scores*. For $log\,odd\,score > 0$ we accept the model, for $log\,odd\,score < 0$ we reject it. Log odd scores have the advantage that we can sum them (instead of multiplying likelihoods) to receive a statistical

meaningful number: That is, the log odd score that all the matched peaks of the measured spectrum belong to their reference counterparts, vs. all the measured peaks are noise.

Now, assume that the model is true, that is, the measured peak belongs to the matched reference peak. Then, it is usually impossible to predict the intensity of the fragment peak solely from its molecular formula. But we can use the mass difference between the measured peak and the molecular formula to assess whether the model holds: We want to assess the likelihood that the mass differences between measured and reference peaks, corresponding to the *mass error* of the measurement, can get this large or larger by chance.

The probability to observe a certain mass error, clearly depends on the accuracy of the instrument: If the instrument has a bad mass accuracy (for example, ion trap MS) than we will observe large mass errors much more often than for an instrument with excellent mass accuracy, such as orbitrap MS. In fact, mass spectrometry literature reports *mass accuracies* of instruments and measurements: This is a unit-free number, usually given in *parts per million* (ppm), showing the relative mass accuracy of the measurement or instrument. For example, if we measure an ion with mass 1000 Da at mass 1000.03 Da, then the "mass accuracy" of the measurement is

$$\frac{|1000 - 1000.03|}{1000} = 3 \cdot 10^{-5} = 30\,\text{ppm}. \tag{4.4}$$

Unfortunately, the mass accuracy reported in the literature often refers to such a single mass difference: Zubarev and Mann [248] coined the term *anecdotal mass accuracy* for the "selective reporting of mass measurements, usually to demonstrate the capabilities of the author's instrument." Such anecdotal mass accuracy "should clearly be distinguished from routine instrument performance in day to day use." Zubarev and Mann also proposed to use the term "*mass deviation*" instead of "mass accuracy" for an individual mass error, such as the one in (4.4).

We need, in contrast, a statistical mass accuracy that assigns probabilities to different mass errors. It turns out that mass errors are roughly normally distributed with mean zero. We can argue statistically, that some random variable that is the sum of numerous other random variables that account to the final peak mass measured in the spectrum, should be normally distributed. But this fact has also been verified experimentally in at least two studies [117, 248].

But before we continue, a *warning* is in place: Observed mass errors are in fact the sum of a systematic mass error due to poor calibration, and the statistical mass introduced above. The systematic mass error can be countered by calibration using (internal or external) standards, or by hypothesis-driven recalibration, see Sec. 7.5 below for more details. Only after we have removed the systematic mass error from the measurement, it is reasonable to assume that mass errors are statistically distributed.

[ToDo: Continue:] and we calculate this likelihood as the two-sided area under the Gaussian curve with SD 1/3 of the relative mass error.

How to define a expedient scoring function? We know from mass spectrometry that mass deviations are nearly uniformly distributed.

The probability $\mathbb{P}(\text{mass deviation} \leq \varepsilon)$ is the integral from both sites, labeled red in Figure 4.1. The expectation value is $\mu = 0$ for a well calibrated instrument. The standard deviation is $\sigma = \frac{1}{3}\frac{z}{10^6}m$, where z is the mass accuracy of the instrument z ppm. We assume that 99,7% of the measurements lie in these area, that is consistent with $\sigma = \frac{1}{3}\frac{z}{10^6}m$. Using the Normal distribution, we cannot rule out that arbitrarily large mass deviations occur; we just assume that they are arbitrarily unlikely. For example, we implicitly assume that 99.9999998% of all measurements are at least within twice the stated mass accuracy, and less than two in a billion

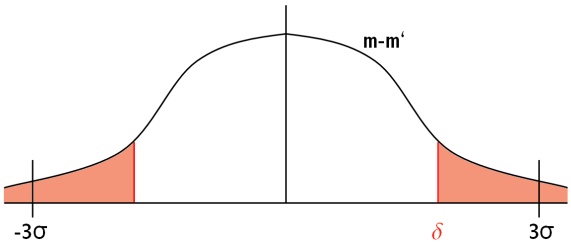

Figure 4.1: We model the distribution of mass deviation ε as a Normal distribution $\mathcal{N}(0,\sigma)$ with mean zero. Then, 99.7 % of the measured mass deviations lie between -3σ and $+3\sigma$.

measurements show a larger mass deviation.[1] We estimate the probability that of observing a mass difference of $|m-m'|$ or larger as:

$$\mathbb{P}(D|H_1) = \mathbb{P}(\text{mass difference of } m-m' \text{ or more}) = \operatorname{erfc}\left(\frac{|m-m'|}{\sqrt{2}\sigma_{\text{mass}}}\right) = \frac{2}{\sqrt{2\pi}}\int_z^\infty e^{-t^2/2}dt \tag{4.5}$$

with $z := \frac{|m-m'|}{\sigma_{\text{mass}}}$, where m, m' are the masses of the measured and the reference peak, and σ is the standard deviation of the Gaussian mass error distribution. **[ToDo: plus c??]** **[ToDo: Explain** erfc**]**

For the background model, we cannot use the mass of the peak since, in general, noise peaks may appear at any mass. But we can use the peak intensity for this purpose: Evaluations have shown that noise peak intensities are roughly exponentially distributed; see for instance Fig. 4 in Goldberg *et al.* [93]. Let $\lambda e^{\lambda x}$ be the exponential distribution with parameter λ, where x is the peak intensity. The likelihood of observing a noise peak with intensity y or higher is

$$\mathbb{P}(\text{intensity noise} \geq y) = \int_y^\infty \lambda e^{\lambda x}dx = e^{-\lambda y}. \tag{4.6}$$

Taking the natural logarithm, we reach $-\lambda y$ for intensity y.

Since this likelihood appears in the denominator of the log odds term, we simply add the peak intensity, multiplied by a constant representing the noise in the spectrum, to the score. Finally, we can use prior probabilities, computing the odds ratio that any peak is not noise: We add a constant b, being the logarithm of this odds ratio, to each vertex score.

It is possible to integrate more peak attributes to the scoring function. For instance high peaks get a better score by adding the intensity to the score, that solves the threshold problem.

To get log odds we assign for each pair of peaks m' and m

$$score(m,m') = \log\frac{\mathbb{P}(\text{peak } m' \text{ is signal at } m)}{\mathbb{P}(\text{peak } m' \text{ is noise peak})} \tag{4.7}$$

If peak m' is signal at m then the intensity of m' is only relevant if we know the intensity of the reference peak m, what we normally don't do. So the mass deviation is $|m-m'|$ If m' is a noise peak there is no mass deviation, so we need a model for the distribution of the intensities of noise peaks. The exponential distribution fits well. For $X \sim Exp(\lambda)$ holds

$$\mathbb{P}(X > x) = e^{-\lambda x} \quad \text{and} \quad \log\mathbb{P}(X > x) = -\lambda x \tag{4.8}$$

[1] We cannot rule out that a meteor hits and destroys the earth tomorrow; it is just very, very unlikely.

The score is know simply the sum of individual scores for all peaks in the spectrum.

Score additional and missing peaks.

We can also take into consideration the mass error of the parent mass, corresponding to the precursor ion.

4.3 Additional peaks: To penalize or not to penalize?

We

The same argumentation holds for scoring the mass error of the parent mass: For a single measured spectrum, this is not required to find the best match in the database. But as soon as we want to compare the score of Peptide-Spectrum Matches for many measured spectra, this score modification should be taken into account.

4.4 Ion series revisited: The abc and xyz of peptide fragmentation

If you take a look at any peptide fragmentation mass spectrum, there is obviously more going on than what we pretended in Chapter 2. Compare Fig. 4.2 below, to Fig. 2.1 on page 16: Besides the two "main" ion series b and y considered so far, there are at least four more ion series, namely a, c, x, and z ions. The following description is tailored towards CID (Collision-Induced Dissociation) peptide fragmentation, which still is the predominant method for this purpose. Keep in mind that there are many obvious and subtle differences for, say, ETD (Electron-Transfer Dissociation) peptide fragmentation.

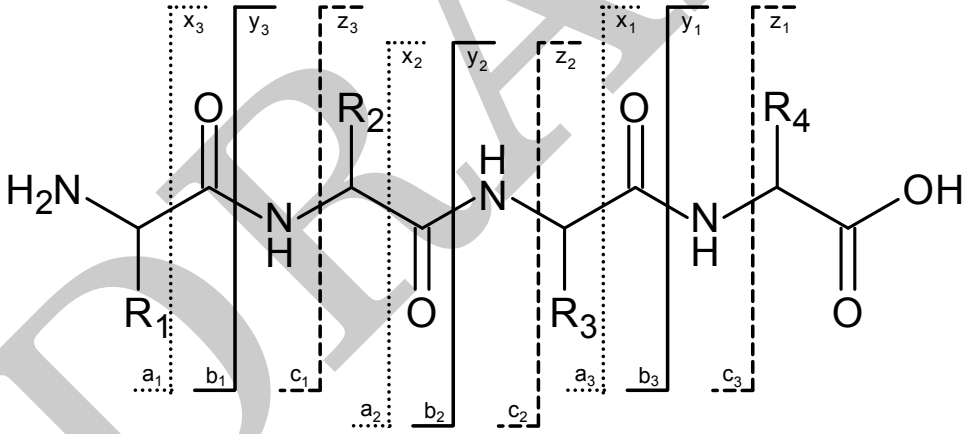

Figure 4.2: Fragmentation of a peptide into a,b,c and x,y,z-ions. This figure oversimplifies the fragmentation process, as hydrogen atom rearrangements are omitted.

But peptide fragmentation is not as simple as Fig. 4.2 might suggest: In fact, fragmentation is a rather involved process that can comprise a series of rearrangements of the molecule, before the actual fragmentation takes place. In fact, peptide fragmentation by Collision-Induced Dissociation is still an active field of research in the mass spectrometry community, even after more than 20 years. We will not go into the details, but rather stick to a mere description of what we find in the spectrum.

prefixes, N-terminal			suffixes, C-terminal		
series	MFM	mass	series	MFM	mass
a ions	$-CO$	-28.000000	x ions	$+CO_2$	$+44.0$
b ions	none	± 0	y ions	$+H_2O$	$+18.0$
c ions	$+NH_3$	$+17.0$	z ions	$-NH_3+H_2O$	$+1.0$
b$-H_2O$, b°	$-H_2O$	-18.0	y$-H_2O$, y°	none	± 0
b$-NH_3$, b*	$-NH_3$	-17.0	y$-NH_3$, y*	$-NH_3+H_2O$	$+1.0$

Table 4.1: The ten most important ion series for peptide fragmentation. 'MFM' is the molecular formula modification, that has to be applied to the molecular formula of the prefix or suffix (without water) to receive the molecular formula of the ion series. As we shift all masses by a proton mass, this mass modification is deliberately excluded from the table. For protonated ions, add H^+ to the molecular formulas, and 1.00728 Da to the mass. **[ToDo: Check MFM, stimmen die jetzt so? Massen ausrechnen]**

The six most important ion series, plus two important modifications are shown in Table 4.1. The a, b, and c ion series correspond to prefixes of the peptide; the x, y, and z ion series correspond to suffixes. To calculate the molecular formula of an ion from a prefix or suffix string, calculate the molecular formula of the residue string using Table 2.1 on page 18; then, add or subtract the molecular formula modification[2] from Table 4.1, plus H^+. We have omitted the proton from the molecular formula modification in Table 4.1 as we will decrease all masses in the measured spectrum by a proton mass, as described in Sec. 2.5.5. Just like b and y ions which are complementary, the same holds for a and x ions, and for c and z ions. The molecular formula of c and z ions adds up to the molecular formula of the peptide. This is not the case for b and y ions, whose molecular formula add up to the peptide molecular minus H_2.

Example 4.4. Given the peptide ESI with molecular formula $C_{14}H_{25}N_3O_7$, assume that ES is the prefix (N-terminal fragment) and I is the suffix (C-terminal fragment). The molecular formula of the residue string ES is $C_5H_7N_1O_3 + C_3H_5N_1O_2 = C_8H_{12}N_2O_5$, whereas the residue string I has molecular formula $C_6H_{11}N_1O_1$. We calculate the corresponding molecular formulas of the ion series as:

series	molecular formula calculation		ion
a ion	$C_8H_{12}N_2O_5 - CO$	$= C_7H_{12}N_2O_4$	$C_7H_{12}N_2O_4H^+$
b ion	$C_8H_{12}N_2O_5$	$= C_8H_{12}N_2O_5$	$C_8H_{12}N_2O_5H^+$
c ion	$C_8H_{12}N_2O_5 + NH_3$	$= C_8H_{15}N_3O_5$	$C_8H_{15}N_3O_5H^+$
x ion	$C_6H_{11}N_1O_1 + CO_2$	$= C_7H_{11}N_1O_3$	$C_7H_{11}N_1O_3H^+$
y ion	$C_6H_{11}N_1O_1 + H_2O$	$= C_6H_{13}N_1O_2$	$C_6H_{13}N_1O_2H^+$
z ion	$C_6H_{11}N_1O_1 - NH_3 + H_2O$	$= C_6H_{10}O_2$	$C_6H_{10}O_2H^+$

Again, the molecular formula calculation is missing the proton for convenience. Now, molecular formulas for a and x ion add up to $C_7H_{12}N_2O_4 + C_7H_{11}N_1O_3 = C_{14}H_{23}N_3O_7$ whereas those for c and z ion add up to $C_8H_{15}N_3O_5 + C_6H_{10}O_2 = C_{14}H_{25}N_3O_7$, the molecular formula of the peptide.

[2] I hereby apologize to all chemists who got a heart attack from seeing the "negative molecular formula," but it really makes things easier. It appears that nobody is publishing these modifications, possibly as you can easily be beaten to death for that by a chemist reviewer. I am taking this opportunity to publish them, once and for all.

Ions of all ion series may also loose ammonia NH_3 or water H_2O. This loss will not happen at the fragmentation side, but somewhere else. In Table 4.1, we show the resulting molecular formula modifications for b and y ions, which are usually the most intense in a CID fragmentation spectrum. It is understood how to calculate molecular formula modifications and masses for the other ion series. Note that z ions and y$-NH_3$ ions have the same molecular formula modifications; in practice, this means that the corresponding peaks are completely indistinguishable, and intensities of the peaks will add up in the mass spectrum.

There can be other significant peaks in the spectrum, that we might want to take into account:

- Immonium ions are produced as a secondary fragmentation of the amide bond, combining an y-type and an a-type fragmentation. From the computational side, this means that we do not see a substring of the peptide string, but rather a single character. Immonium ions have structural formula $[H_3N{=}CH{-}R]^+$ where R is the amino acid side chain. The molecular formula of an immonium ion is the molecular formula of the amino acid residue, plus CH_3N (without charge) or plus CH_3NH^+ (for the ion). Every amino acid has a corresponding immonium ion, but fragments of immonium ions can also be observed. Immonium ions cannot be used for determining the sequence of the peptide, but they are indicative of the presence or absence of a particular amino acid from the sequence.
- The spectrum may contain multiple-charged molecules, see Chapter 7. Note that doubly-charged fragments can be common in tandem mass spectra, as the precursor ion is multiple charged.
- Internal ions are substrings of s that are neither suffixes nor prefixes; internal ions have terminal composition $+H_3O$ and are usually rare in tandem mass spectra.

4.5 Even smarter scoring: Know your application

In Sec. 4.2 we have used several factors to score the measured spectrum against a reference: Namely, the mass error of fragment ions; peak intensities; missing and additional peaks; the mass error of the parent mass; and, by summing up all individual scores, also the number of matched peaks. Now, let us assume that our measured spectrum stems from CID peptide fragmentation. Then, we can use a much more sophisticated score that, in particular, takes into account the *dependencies* between different ion series.

But firstly, we can use *immonium ions* to modify the score: these ions are indicative of the presence or absence of a particular amino acid from the sequence. So, if our candidate peptide contains some amino acid x, then we can reward the existence and penalize the non-existence of the corresponding immonium ion peak. If, in contrast, our candidate peptide does not contain some amino acid x,

The following list is tailored towards CID (Collision-Induced Dissociation) peptide fragmentation, and a much different list is needed for, say, ETD (Electron-Transfer Dissociation) peptide fragmentation. The important point here, is that *some* dependencies exist, and can be used to improve the score.

- If both the b and y ion are present in the measured spectrum, this is a better indication than a single b or y ion, even if the intensity of the single peak is higher than the summed intensities of both b and y ion.
- A similar argument holds for a y ion with a water or ammonia loss.

- Y ions tend to appear in consecutive series, so a "ladder" of five y ions should be scored better than five y ions distributed with gaps across the peptide sequence.
- All of the above can be used to modify our score, depending on the intensities of the corresponding peaks.

4.6 Peptide database search programs

See Aaant [1], Steen and Mann [218] or many other reviews for searching peptides in databases. The following is a list of available approaches for searching peptide databases; I expect it to be incomplete. The most important database search programs are still MASCOT (which is based on MOWSE), SEQUEST, and X!Tandem. The other tools are listed in (mostly) chronological order.

MOWSE (MOlecular Weight SEarch) was developed in 1993 by Pappin *et al*. [182] and was initially targeted at the identification of proteins using peptide mass fingerprints. The reference spectra for each entry in the sequence database are calculated in the preprocessing for a faster search.

Uses average properties of the proteins in the database to improve the sensitivity and selectivity of the identification. It takes into account the relative abundance of the peptides in the database when calculating the score, that is, the chance of getting a random match to a larger peptide is lower and therefore it will contribute to a higher degree to the score. Also the protein size effect is compensated for.

MASCOT by Perkins *et al*. [186] is the successor of MOWSE and is available at www.matrixscience.com. It can be used for peptide-mass finger print and MS/MS. No preprocessing of the database is needed. MASCOT converts the MOWSE score to a probability, that the score was achieved accidentally.

$$MASCOTScore = -10 \log_{10} p \tag{4.9}$$

The heuristic which is not documented performs well. It is based on a client/server model, the central MASCOT server searches the database. It handles peptide modifications etc pp. The performance on MS/MS got better.

SEQUEST by Eng *et al*. [70] is one of the first tools for searching peptide fragmentation spectra in databases, see Sadygov *et al*. [200] some details. SEQUEST is commercially available from Thermo Finnigan. The software proceeds in two passes: In a first pass, a very simple scoring is used to find 500 candidates that match the measured spectrum reasonable well. In the second pass, the reference spectrum is simulated as a "raw" spectrum. Here, the score of a simulated reference spectrum $f : \mathbb{R} \to \mathbb{R}$ and a measured spectrum $g : \mathbb{R} \to \mathbb{R}$ is computed by correlation as $\text{score}(f,g) = \int_{m \in \mathbb{R}} f(m) \cdot g(m)\, dm$. SEQUEST sorts the 500 candidates according to this score. SEQUEST reaches good search results, but the software is very slow, attributed to it processing "raw" spectral data instead of peak lists. In fact, it does not compute a single convolution but instead, uses the Fourier transform to compute several such scores for different mass shifts of the measured spectrum, that are used to normalize the final score. This is even more time consuming. As we will see in Sec. 4.7, the correlation approach of SEQUEST is not very far from what we have done in

Sec. 4.2; this is possibly the explanation why SEQUEST performs very good in practice. On the downside, the correlation approach requires much more time, and has not statistical justification.

X!Tandem by Craig and Beavis [47] is an open source tool available from Global Proteome Machine Organization (`http://www.thegpm.org`). It is used for MS/MS and database search and assigns a simple scoring and calculates expectation values and significances. The latter can be also calculated for protein identifications. X!!Tandem is the parallelized version. Other Tandem MS search engines are X!P3 and X!Hunter.

VEMS was developed by Matthiesen et al in 2005, the latest release is VEMS 3.0. It is a good alternative to MASCOT. It is not based on new ideas, but the most ideas and solutions are included.

xxx by Elias *et al*. [69]

xxx by McHugh and Arthur [165]

PepProbe by Sadygov and Yates III [199]

ProteinProspector by Clauser *et al*. [44]

Crux by Park *et al*. [183]

MS-Dictionary by Kim *et al*. [135] is somewhat different, as it combines *de novo* sequencing and database searching. **[TODO: MAYBE, GIVE IT A SEPARATE SECTION? HOW DOES IT DIFFER FROM KIM *et al*. [134]??]**

Evaluations of these database search tools were performed by Chamrad *et al*. [36]. **[TODO: LOOK AT KAPP *et al*. [126]]**

There are also tools that do a kind of "database post-processing":

Peptide Prophet by [128] tries to estimate the probability that the identification is correct, using the score of any search engine. This is achieved by empirically estimating a score distribution. It is able to combine the outputs of several search engines to a single score.

PERCOLATOR by Käll *et al*. [123] Extract multiple features from PSMs reported by other database search tools. Provide the target and decoy database (a randomized protein database) to the tool, see Chapter 5. Iteratively adjust weights of features to maximize the number of peptides matched to the target database at a target false discovery rate (FDR) using a Support Vector Machine (SVM). It greatly improves the number of Peptide-Spectrum Matches; but this comes at the cost that we might get results that *look* like what we expect; but these might be arbitrarily far from the truth [38].

4.7 The maths of SEQUEST scoring

Not every detail of the original SEQUEST scoring has been described, and some things might remain mysterious forever, as SEQUEST is now commercial, closed-source software. Still and all, the idea of *correlating* the theoretical and measured spectra keeps reappearing in the computational MS literature, too. This is motivation enough for us to take a short look at

the maths behind this type of scoring, and to explain why it is ill-chosen for searching peptide tandem mass spectra.

Now, assume that the theoretical spectrum is given as a function $f : \mathbb{R} \to \mathbb{R}$ and, similarly, the measured spectrum is another function $g : \mathbb{R} \to \mathbb{R}$. In reality, both spectra are discrete lists of measurements, but for our mathematical considerations it is easier to think of them as continuous function. For ease of presentation, let us concentrate on a single peak in the theoretical spectrum, to be matched with a single peak in the measured spectrum. Assume that the peak masses differ by $\delta \in \mathbb{R}$. Without loss of generality, let the peak mass in the theoretical spectrum be zero. It is not fully described how SEQUEST simulates peaks in the theoretical spectrum, but as the authors talk about "peak width" [70], it is probably not "sticks" (Dirac's delta function). For simplicity, we assume that they model the peak shape as a Gaussian function,

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-x^2}{2\sigma^2}\right).$$

Even though calculations for MS physics tell us that peaks should have shapes slightly different from a Gaussian function, it is still common to model them as a Gaussian, as these are easier to "handle" from the numerical standpoint. Furthermore, assume that a peak in the measured spectrum has the ideal shape of a Gaussian function, so $g(x) = f(x-\delta)$. If the measured peak differs from the ideal model shape, this means that the score is reduced. Both peaks have identical width (parameter variance σ^2), but this is done solely to simplify our calculations below.

SEQUEST (following other publications who described this idea earlier) uses a convolution $\int f(x)g(x)\,dx$ to compute the score for the theoretical spectrum, given the measured spectrum.

$$\begin{aligned}
\int f(x)\cdot g(x)\,dx &= \int f(x)\cdot f(x-\delta)\,dx \\
&= \int \frac{1}{2\pi\sigma^2} \exp\left(\frac{-x^2-(x-\delta)^2}{2\sigma^2}\right) dx \\
&= \frac{1}{2\pi\sigma^2} \int \exp\left(\frac{-2x^2+2\delta x-\delta^2}{2\sigma^2}\right) dx \\
&= \frac{1}{2\pi\sigma^2} \int \exp\left(\frac{-x^2+\delta x-\frac{1}{2}\delta^2}{\sigma^2}\right) dx \\
&= \frac{1}{2\pi\sigma^2} \int \exp\left(\frac{-(x-\frac{1}{2}\delta)^2+\frac{1}{4}\delta^2-\frac{1}{2}\delta^2}{\sigma^2}\right) dx \\
&= \frac{1}{2\pi\sigma^2} \cdot \exp\left(\frac{-\frac{1}{4}\delta^2}{\sigma^2}\right) \cdot \int \exp\left(\frac{-(x-\frac{1}{2}\delta)^2}{\sigma^2}\right) dx \\
&= \frac{\sqrt{\pi\sigma^2}}{2\pi\sigma^2} \cdot \exp\left(\frac{-\delta^2}{4\sigma^2}\right) \cdot \frac{1}{\sqrt{2\pi(\sigma/\sqrt{2})^2}} \int \exp\left(\frac{-(x-\frac{1}{2}\delta)^2}{2(\sigma/\sqrt{2})^2}\right) dx
\end{aligned} \tag{4.10}$$

For the definite integral we calculate

$$\int_a^b f(x)\cdot g(x)\,dx = \frac{1}{2\sqrt{\pi\sigma^2}} \cdot \exp\left(\frac{-\delta^2}{4\sigma^2}\right) \cdot \left[F\left(\tfrac{b-\delta/2}{\sigma/\sqrt{2}}\right) - F\left(\tfrac{a-\delta/2}{\sigma/\sqrt{2}}\right)\right] \tag{4.11}$$

where $F(x) := \frac{1}{2} + \frac{1}{2}\operatorname{erf} x$ is the cumulative distribution function of the normal distribution, and "erf" is the error function, see above.

For the integral from $a = -\infty$ to $b = +\infty$ we reach

$$\text{score}(f,g) = \int_{-\infty}^{+\infty} f(x)g(x)\,dx = \frac{1}{\sqrt{2\pi \cdot 2\sigma^2}} \cdot \exp\left(-\frac{\delta^2}{2 \cdot 2\sigma^2}\right) \tag{4.12}$$

which is just the Normal distribution with mean 0 and variance $2\sigma^2$. Compared to Sec. 4.2, we have replaced the probability that we see a mass deviation of at least δ under the normal distribution, by the relative likelihood that exactly this mass deviation occurs, also under the normal distribution. We have mentioned above that this is not a probability: The probability for mass deviation exactly δ is zero, for any δ.

As function g corresponding to the measured spectrum is in fact recorded, say, as integer values between 0 and 1023, we may assume that g has finite support: That is, we assume **[ToDo: Pass op!]**

This is what it boils down to: The correlation scoring is very time consuming to compute. We sacrifice our statistical model of mass deviation, but still end up with a score similar to the probability density function of a Normal distribution, which has no statistical interpretation. Novel methods for peak picking have surpassed the simple correlation of peak shapes for a long time, see Sec. 15.2. Peak shape scores can be easily incorporated into our scoring from Sec. 4.2; the high running times of the correlation approach can be explained by evaluating peak shapes over and over again. For the application at hand, I do not see any advantages of the correlation approach:[3] Computing such correlation is useful when one has no model whatsoever about the data at hand, such as comparing *two measured* spectra. But if one first has to simulate one of the spectra, then computing correlations is an unnecessary and misleading detour.

4.8 Historical notes and further reading

The title of Sec. 4.4 was borrowed from the paper by Steen and Mann [218]. The nomenclature of ion series is due to Roepstorff and Fohlman [197].

Zubarev and Mann [248] propose to use known peptides as internal calibrants, until the distribution of mass errors is normally distributed. The paper also contains some details on mass accuracy needed to identify peptides an proteins from their monoisotopic mass; we will come back to this in Sec. 10.8.

Using peak intensities to score the peaks in a spectrum, as explained in Sec. 4.2, has been proposed many times in the literature [1, 157], but this is usually done without giving any (stochastic) justification. **[ToDo: Check Havilio *et al.* [108].]** Goldberg *et al.* [93] suggested to use $\exp(a_0 + a_1 x + a_2 x^2)$ to model intensities of noise peaks; but it seems that a_0 is *very* close to zero (see Fig. 4, right in their paper), so that we are back to an exponential distribution. Also, their model has the not appealing property that noise peaks with negative intensity have probabilities larger than zero. The fact that the distribution is truncated for low intensities, can be attributed to thresholding in the peak picking algorithm.

Designing good scoring functions for scoring peptide fragmentation mass spectra was, is, and will be an area of active research [7, 44, 48, 70, 79, 108, 135, 156, 160, 163, 164, 170, 186, 210, 221, 223]. **[ToDo: Was schlage ich vor?]**

Matthiesen *et al.* [164] describe how to score the presence or absence of immonium ions. Different from their Table 1, I would rather suggest to use Machine Learning to estimate the

[3] With the possible exception that it allows computer manufacturers to sell more of their expensive compute clusters.

required probabilities, as this could take into account dependencies between the fragments of an immonium ion.

Back in 1992, Owens [179] suggested to use correlation of mass spectra as a score.

4.9 Exercises

4.1 Write an algorithm to simulate the tandem MS spectrum of a peptide if only b and y ions are present, and ions have a single proton.

4.2 Given two peak lists $\mathcal{M} = \{150, 175, 220, 310, 470\}$ and $\mathcal{M}' = \{150, 190, 250, 315, 485\}$. Calculate the peak counting score for $\varepsilon = 5$ and for $\varepsilon = 15$.

4.3 Let $\text{score}(i,j) = 2 - \frac{1}{5}\left|m_i - m'_j\right|$ and $\text{score}(i,\epsilon) = \text{score}(\epsilon,j) = -1$. Calculate an optimal alignment of these two peak lists.

4.4 Given a tandem mass spectrum

$$\mathcal{M} = \{440, 682, 748, 753, 837, 870, 884, 1017, 1196, 1235, 2637\}$$

with masses in Dalton. **[TODO: MAYBE, DIFFERENT LIST OF MASSES?]** Use the MASCOT web interface at `http://www.matrixscience.com` to find the protein that this peptide might come from. As option, you may assume that trypsin has been used for digestion, that there are no modifications, that average masses were recorded, and that the mass accuracy is **[TODO: SOMETHING]**. **[TODO: WHY AVERAGE MASSES??]** Use SwissProt as you protein database. Look at that *Protein Summary* in the output: What is the meaning of columns *Observed*, *Mr(expt)*, *Delta*, *Start*, *End*, emphMiss, and *Peptide*?

4.5 **log-Scores** Viele Programme (z.B. Mascot und SCOPE) berechnen Wahrscheinlichkeiten als Scores. Von diesen Programmen wird aber meist nicht der Score s zurückgegeben, sondern $-\log s$. Welche Vorteile hat das?

4.6 Why is it reasonable, from a mathematical perspective, to assume that for 10 ppm mass accuracy, the difference between two peak masses has at most 14 ppm? Hint: if a random variable X is normally distributed, than $-X$ is so, too. Under what condition is it reasonable, from a MS perspective, to assume that this mass difference is still 10 ppm?

4.7 Write an algorithm to simulate the tandem MS spectrum of a peptide as thoroughly as possible.

4.8 Proof Lemma 4.1.

Bibliography

[1] A. Aaant. I need a title, quick. **[TODO: REPLACE WITH A REAL CITATION]**, 2101.

[2] G. Alves, A. Y. Ogurtsov and Y.-K. Yu. RAId_DbS: peptide identification using database searches with realistic statistics. *Biol. Direct.*, 2:25, 2007.

[3] S. Andreotti, G. W. Klau and K. Reinert. Antilope – a lagrangian relaxation approach to the *de novo* peptide sequencing problem. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 2011. To appear, doi:10.1109/TCBB.2011.59.

[4] R. Apweiler, H. Hermjakob and N. Sharon. On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochim. Biophys. Acta*, 1473(1): 4–8, 1999.

[5] G. Audi, A. Wapstra and C. Thibault. The AME2003 atomic mass evaluation (ii): Tables, graphs, and references. *Nucl. Phys. A*, 729:129–336, 2003.

[6] J.-M. Autebert, J. Berstel and L. Boasson. Context-free languages and pushdown automata. In G. Rozenberg and A. Salomaa, editors, *Handbook of Formal Languages*, volume 1, pages 111–174. Springer, 1997.

[7] V. Bafna and N. Edwards. SCOPE: A probabilistic model for scoring tandem mass spectra against a peptide database. *Bioinformatics*, 17:S13–S21, 2001.

[8] D. A. Barkauskas and D. M. Rocke. A general-purpose baseline estimation algorithm for spectroscopic data. *Anal. Chim. Acta*, 657(2):191–197, 2010.

[9] C. Bartels. Fast algorithm for peptide sequencing by mass spectrometry. *Biomed. Environ. Mass Spectrom.*, 19:363–368, 1990.

[10] J. M. S. Bartlett and D. Stirling. A short history of the polymerase chain reaction. *Methods Mol. Biol.*, 226:3–6, 2003.

[11] C. Bauer, R. Cramer and J. Schuchhardt. Evaluation of peak-picking algorithms for protein mass spectrometry. *Methods Mol. Biol.*, 696:341–352, 2011.

[12] M. Beck, I. M. Gessel and T. Komatsu. The polynomial part of a restricted partition function related to the Frobenius problem. *Electron. J. Comb.*, 8(1):N7, 2001.

[13] D. E. Beihoffer, J. Hendry, A. Nijenhuis and S. Wagon. Faster algorithms for Frobenius numbers. *Electron. J. Comb.*, 12:R27, 2005.

[14] C. Benecke, T. Grüner, A. Kerber, R. Laue and T. Wieland. MOLecular Structure GENeration with MOLGEN, new features and future developments. *Anal. Chim. Acta*, 314:141–147, 1995.

[15] G. Benson. Composition alignment. In *Proc. of Workshop on Algorithms in Bioinformatics (WABI 2003)*, volume 2812 of *Lect. Notes Comput. Sc.*, pages 447–461. Springer, 2003.

[16] M. W. Bern and D. Goldberg. EigenMS: De novo analysis of peptide tandem mass spectra by spectral graph partitioning. In *Proc. of Research in Computational Molecular Biology (RECOMB 2005)*, volume 3500 of *Lect. Notes Comput. Sc.*, pages 357–372. Springer, 2005.

[17] M. W. Bern and D. Goldberg. De novo analysis of peptide tandem mass spectra by spectral graph partitioning. *J. Comput. Biol.*, 13(2):364–378, 2006.

[18] A. Bertsch, A. Leinenbach, A. Pervukhin, M. Lubeck, R. Hartmer, C. Baessmann, Y. A. Elnakady, R. Müller, S. Böcker, C. G. Huber, and O. Kohlbacher. De novo peptide sequencing by tandem MS using complementary CID and electron transfer dissociation. *Electrophoresis*, 30(21):3736–3747, 2009.

[19] K. Biemann, C. Cone and B. R. Webster. Computer-aided interpretation of high-resolution mass spectra. II. Amino acid sequence of peptides. *J. Am. Chem. Soc.*, 88(11):2597–2598, 1966.

[20] K. Biemann, C. Cone, B. R. Webster and G. P. Arsenault. Determination of the amino acid sequence in oligopeptides by computer interpretation of their high-resolution mass spectra. *J. Am. Chem. Soc.*, 88(23):5598–5606, 1966.

[21] A. Björklund, T. Husfeldt, P. Kaski and M. Koivisto. Fourier meets Möbius: fast subset convolution. In *Proc. of ACM Symposium on Theory of Computing (STOC 2007)*, pages 67–74. ACM Press New York, 2007.

[22] N. Blow. Glycobiology: A spoonful of sugar. *Nature*, 457(7229):617–620, 2009.

[23] S. Böcker. Sequencing from compomers: Using mass spectrometry for DNA de-novo sequencing of 200+ nt. *J. Comput. Biol.*, 11(6):1110–1134, 2004.

[24] S. Böcker and Zs. Lipták. A fast and simple algorithm for the Money Changing Problem. *Algorithmica*, 48(4):413–432, 2007.

[25] S. Böcker and V. Mäkinen. Combinatorial approaches for mass spectra recalibration. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 5(1):91–100, 2008.

[26] S. Böcker and F. Rasche. Towards de novo identification of metabolites by analyzing tandem mass spectra. *Bioinformatics*, 24:I49–I55, 2008. Proc. of *European Conference on Computational Biology* (ECCB 2008).

[27] S. Böcker, M. Letzel, Zs. Lipták and A. Pervukhin. Decomposing metabolomic isotope patterns. In *Proc. of Workshop on Algorithms in Bioinformatics (WABI 2006)*, volume 4175 of *Lect. Notes Comput. Sc.*, pages 12–23. Springer, 2006.

[28] S. Böcker, B. Kehr and F. Rasche. Determination of glycan structure from tandem mass spectra. In *Proc. of Computing and Combinatorics Conference (COCOON 2009)*, volume 5609 of *Lect. Notes Comput. Sc.*, pages 258–267. Springer, 2009.

[29] S. Böcker, M. Letzel, Zs. Lipták and A. Pervukhin. SIRIUS: Decomposing isotope patterns for metabolite identification. *Bioinformatics*, 25(2):218–224, 2009.

[30] S. Böcker, F. Rasche and T. Steijger. Annotating fragmentation patterns. In *Proc. of Workshop on Algorithms in Bioinformatics (WABI 2009)*, volume 5724 of *Lect. Notes Comput. Sc.*, pages 13–24. Springer, 2009.

[31] A. Brauer and J. E. Shockley. On a problem of Frobenius. *J. Reine Angew. Math.*, 211: 215–220, 1962.

[32] R. Breitling, A. R. Pitt and M. P. Barrett. Precision mapping of the metabolome. *Trends Biotechnol.*, 24(12):543–548, 2006.

[33] K. Q. Brown. *Geometric transforms for fast geometric algorithms*. Report cmucs-80-101, Dept. Comput. Sci., Carnegie-Mellon Univ., Pittsburgh, USA, 1980.

[34] S. Cappadona, P. Nanni, M. Benevento, F. Levander, P. Versura, A. Roda, S. Cerutti, and L. Pattini. Improved label-free LC-MS analysis by wavelet-based noise rejection. *J Biomed Biotechnol*, 2010:131505, 2010.

[35] A. Ceroni, K. Maass, H. Geyer, R. Geyer, A. Dell and S. M. Haslam. GlycoWorkbench: a tool for the computer-assisted annotation of mass spectra of glycans. *J. Proteome Res.*, 7 (4):1650–1659, 2008.

[36] D. C. Chamrad, G. Körting, K. Stühler, H. E. Meyer, J. Klose and M. Blüggel. Evaluation of algorithms for protein identification from sequence databases using mass spectrometry data. *Proteomics*, 4:619–628, 2004.

[37] S. Chattopadhyay and P. Das. The K-dense corridor problems. *Pattern Recogn. Lett.*, 11 (7):463–469, 1990.

[38] E. Check. Proteomics and cancer: Running before we can walk? *Nature*, 429:496–497, 2004.

[39] T. Chen, M.-Y. Kao, M. Tepel, J. Rush and G. M. Church. A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry. *J. Comput. Biol.*, 8(3):325–337, 2001. Preliminary version in *Proc. of Symposium on Discrete Algorithms (SODA 2000)*, Association for Computing Machinery, 2000, 389–398.

[40] W. L. Chen. Chemoinformatics: past, present, and future. *J. Chem. Inf. Model.*, 46(6): 2230–2255, 2006.

[41] F. Y. Chin, C. A. Wang and F. L. Wang. Maximum stabbing line in 2D plane. In *Proc. of Conf. on Computing and Combinatorics (COCOON 1999)*, volume 1627 of *Lect. Notes Comput. Sc.*, pages 379–388. Springer, 1999.

[42] H. H. Chou, H. Takematsu, S. Diaz, J. Iber, E. Nickerson, K. L. Wright, E. A. Muchmore, D. L. Nelson, S. T. Warren, and A. Varki. A mutation in human CMP-sialic acid hydroxylase occurred after the Homo-Pan divergence. *Proc. Natl. Acad. Sci. U. S. A.*, 95(20):11751–11756, 1998.

[43] Y. Chu and T. Liu. On the shortest arborescence of a directed graph. *Sci. Sinica*, 14: 1396–1400, 1965.

[44] K. R. Clauser, P. Baker and A. L. Burlingame. Role of accurate mass measurement (+/− 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal. Chem.*, 71(14):2871–2882, 1999.

[45] C. A. Cooper, E. Gasteiger and N. H. Packer. GlycoMod – a software tool for determining glycosylation compositions from mass spectrometric data. *Proteomics*, 1(2):340–349, 2001.

[46] C. A. Cooper, H. J. Joshi, M. J. Harrison, M. R. Wilkins and N. H. Packer. GlycoSuiteDB: a curated relational database of glycoprotein glycan structures and their biological sources. 2003 update. *Nucleic Acids Res.*, 31(1):511–513, 2003.

[47] R. Craig and R. C. Beavis. Tandem: matching proteins with tandem mass spectra. *Bioinformatics*, 20(9):1466–1467, 2004.

[48] V. Dančík, T. A. Addona, K. R. Clauser, J. E. Vath and P. A. Pevzner. De novo peptide sequencing via tandem mass spectrometry: A graph-theoretical approach. *J. Comput. Biol.*, 6(3/4):327–342, 1999. Preliminary version in *Proc. of Research in Computational Molecular Biology (RECOMB 1999)*, 135–144.

[49] C. Dass. *Principles and practice of biological mass spectrometry*. John Wiley and Sons, 2001.

[50] R. Datta and M. W. Bern. Spectrum fusion: using multiple mass spectra for de novo peptide sequencing. *J. Comput. Biol.*, 16(8):1169–1182, 2009.

[51] J. L. Davison. On the linear diophantine problem of Frobenius. *J. Number Theory*, 48(3): 353–363, 1994.

[52] M. de Berg, M. van Kreveld, M. Overmars and O. Schwarzkopf. *Computational Geometry: Algorithms and Applications*. Springer, second edition, 2000.

[53] E. de Hoffmann and V. Stroobant. *Mass Spectrometry: Principles and Applications*. Wiley-Interscience, third edition, 2007.

[54] J. R. de Laeter, J. K. Böhlke, P. D. Biévre, H. Hidaka, H. S. Peiser, K. J. R. Rosman and P. D. P. Taylor. Atomic weights of the elements. Review 2000 (IUPAC technical report). *Pure Appl. Chem.*, 75(6):683–800, 2003.

[55] E. W. Deutsch, H. Lam and R. Aebersold. Data analysis and bioinformatics tools for tandem mass spectrometry in proteomics. *Physiological Genomics*, 33:18–25, 2008.

[56] P. A. DiMaggio and C. A. Floudas. De novo peptide identification via tandem mass spectrometry and integer linear optimization. *Anal. Chem.*, 79(4):1433–1446, 2007.

[57] B. Domon and R. Aebersold. Mass spectrometry and protein analysis. *Science*, 312:212–217, 2006.

[58] B. Domon and C. E. Costello. A systematic nomenclature for carbohydrate fragmentations in FAB-MS/MS spectra of glycoconjugates. *Glycoconjugate J.*, 5:397–409, 1988.

[59] R. Dondi, G. Fertin and S. Vialette. Complexity issues in vertex-colored graph pattern matching. *J. Discrete Algorithms*, 2010. In press, doi:10.1016/j.jda.2010.09.002.

[60] R. G. Downey and M. R. Fellows. *Parameterized Complexity*. Springer, 1999.

[61] S. E. Dreyfus and R. A. Wagner. The Steiner problem in graphs. *Networks*, 1(3):195–207, 1972.

[62] M. Dyer. Approximate counting by dynamic programming. In *Proc. of Symposium on Theory of Computing (STOC 2003)*, pages 693–699, 2003.

[63] S. R. Eddy. "antedisciplinary" science. *PLoS Comput. Biol.*, 1(1):e6, 2005.

[64] P. Edman. Method for determination of the amino acid sequence in peptides. *Acta Chem. Scand.*, 4:283–293, 1950.

[65] J. Edmonds. Optimum branchings. *J. Res. Nat. Bur. Stand.*, 71B:233–240, 1967.

[66] M. Ehrich, S. Böcker and D. van den Boom. Multiplexed discovery of sequence polymorphisms using base-specific cleavage and MALDI-TOF MS. *Nucleic Acids Res.*, 33(4):e38, 2005.

[67] D. Einstein, D. Lichtblau, A. Strzebonski and S. Wagon. Frobenius numbers by lattice point enumeration. *INTEGERS*, 7(1):#A15, 2007.

[68] J. E. Elias and S. P. Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods*, 4(3):207–214, 2007.

[69] J. E. Elias, F. D. Gibbons, O. D. King, F. P. Roth and S. P. Gygi. Intensity-based protein identification by machine learning from a library of tandem mass spectra. *Nat. Biotechnol.*, 22(2):214–219, 2004.

[70] J. K. Eng, A. L. McCormack and J. R. Yates III. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectr.*, 5:976–989, 1994.

[71] M. Ethier, J. A. Saba, M. Spearman, O. Krokhin, M. Butler, W. Ens, K. G. Standing, and H. Perreault. Application of the StrOligo algorithm for the automated structure assignment of complex N-linked glycans from glycoproteins using tandem mass spectrometry. *Rapid Commun. Mass Spectrom.*, 17(24):2713–2720, 2003.

[72] M. Fellows, G. Fertin, D. Hermelin and S. Vialette. Sharp tractability borderlines for finding connected motifs in vertex-colored graphs. In *Proc. of International Colloquium on Automata, Languages and Programming (ICALP 2007)*, volume 4596 of *Lect. Notes Comput. Sc.*, pages 340–351. Springer, 2007.

[73] J. Fenn, M. Mann, C. Meng, S. Wong and C. Whitehouse. Electrospray ionisation for mass spectrometry of large biomolecules. *Science*, 246:64–71, 1989.

[74] D. Fenyö and R. C. Beavis. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal. Chem.*, 75(4):768–774, 2003.

[75] J. Fernández-de-Cossío, L. J. Gonzalez and V. Besada. A computer program to aid the sequencing of peptides in collision-activated decomposition experiments. *Comput. Appl. Biosci.*, 11(4):427–434, 1995.

[76] J. Fernández-de-Cossío, J. Gonzalez, T. Takao, Y. Shimonishi, G. Padron and V. Besada. A software program for the rapid sequence analysis of unknown peptides involving modifications, based on MS/MS data. In *ASMS Conf. on Mass Spectrometry and Allied Topics, Slot 074*, 1997.

[77] J. Fernández-de-Cossío, L. J. Gonzalez, Y. Satomi, L. Betancourt, Y. Ramos, V. Huerta, A. Amaro, V. Besada, G. Padron, N. Minamino, and T. Takao. Isotopica: a tool for the calculation and viewing of complex isotopic envelopes. *Nucleic Acids Res.*, 32(Web Server issue):W674–W678, 2004.

[78] A. R. Fernie, R. N. Trethewey, A. J. Krotzky and L. Willmitzer. Metabolite profiling: from diagnostics to systems biology. *Nat. Rev. Mol. Cell Biol.*, 5(9):763–769, 2004.

[79] H. I. Field, D. Fenyö and R. C. Beavis. RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimises protein identification, and archives data in a relational database. *Proteomics*, 2(1):36–47, 2002.

[80] B. Fischer, V. Roth, F. Roos, J. Grossmann, S. Baginsky, P. Widmayer, W. Gruissem, and J. M. Buhmann. NovoHMM: a hidden Markov model for de novo peptide sequencing. *Anal. Chem.*, 77(22):7265–7273, 2005.

[81] P. Flajolet and R. Sedgewick. *Analytic Combinatorics*. Cambridge University Press, 2009. Freely available from `http://algo.inria.fr/flajolet/Publications/book.pdf`.

[82] A. Frank and P. Pevzner. PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal. Chem.*, 15:964–973, 2005.

[83] A. M. Frank, M. M. Savitski, M. N. Nielsen, R. A. Zubarev and P. A. Pevzner. De novo peptide sequencing and identification with precision mass spectrometry. *J. Proteome Res.*, 6(1):114–123, 2007.

[84] A. Fürst, J.-T. Clerc and E. Pretsch. A computer program for the computation of the molecular formula. *Chemom. Intell. Lab. Syst.*, 5:329–334, 1989.

[85] V. A. Fusaro, D. R. Mani, J. P. Mesirov and S. A. Carr. Prediction of high-responding peptides for targeted protein assays by mass spectrometry. *Nat. Biotechnol.*, 27(2):190–198, 2009.

[86] H. Gabow, Z. Galil, T. Spencer and R. Tarjan. Efficient algorithms for finding minimum spanning trees in undirected and directed graphs. *Combinatorica*, 6:109–122, 1986.

[87] M. R. Garey and D. S. Johnson. *Computers and Intractability (A Guide to Theory of NP-Completeness)*. Freeman, New York, 1979.

[88] J. Gasteiger, W. Hanebeck and K.-P. Schulz. Prediction of mass spectra from structural information. *J. Chem. Inf. Comput. Sci.*, 32(4):264–271, 1992.

[89] S. P. Gaucher, J. Morrow and J. A. Leary. STAT: a saccharide topology analysis tool used in combination with tandem mass spectrometry. *Anal. Chem.*, 72(11):2331–2336, 2000.

[90] L. Y. Geer, S. P. Markey, J. A. Kowalak, L. Wagner, M. Xu, D. M. Maynard, X. Yang, W. Shi, and S. H. Bryant. Open mass spectrometry search algorithm. *J. Proteome Res.*, 3:958–964, 2004.

[91] P. Gilmore and R. Gomory. Multi-stage cutting stock problems of two and more dimensions. *Oper. Res.*, 13(1):94–120, 1965.

[92] D. Goldberg, M. Sutton-Smith, J. Paulson and A. Dell. Automatic annotation of matrix-assisted laser desorption/ionization N-glycan spectra. *Proteomics*, 5(4):865–875, 2005.

[93] D. Goldberg, M. W. Bern, B. Li and C. B. Lebrilla. Automatic determination of O-glycan structure from fragmentation spectra. *J. Proteome Res.*, 5(6):1429–1434, 2006.

[94] D. Goldberg, M. W. Bern, S. Parry, M. Sutton-Smith, M. Panico, H. R. Morris and A. Dell. Automated N-glycopeptide identification using a combination of single- and tandem-MS. *J. Proteome Res.*, 6(10):3995–4005, 2007.

[95] D. Goldberg, M. W. Bern, S. J. North, S. M. Haslam and A. Dell. Glycan family analysis for deducing N-glycan topology from single MS. *Bioinformatics*, 25(3):365–371, 2009.

[96] A. H. Grange, M. C. Zumwalt and G. W. Sovocool. Determination of ion and neutral loss compositions and deconvolution of product ion mass spectra using an orthogonal acceleration time-of-flight mass spectrometer and an ion correlation program. *Rapid Commun. Mass Spectrom.*, 20(2):89–102, 2006.

[97] N. A. Gray. Applications of artificial intelligence for organic chemistry: Analysis of C-13 spectra. *Artificial Intelligence*, 22(1):1–21, 1984.

[98] N. A. B. Gray, R. E. Carhart, A. Lavanchy, D. H. Smith, T. Varkony, B. G. Buchanan, W. C. White, and L. Creary. Computerized mass spectrum prediction and ranking. *Anal. Chem.*, 52(7):1095–1102, 1980.

[99] N. A. B. Gray, A. Buchs, D. H. Smith and C. Djerassi. Computer assisted structural interpretation of mass spectral data. *Helv. Chim. Acta*, 64(2):458–470, 1981.

[100] H. Greenberg. Solution to a linear diophantine equation for nonnegative integers. *J. Algorithms*, 9(3):343–353, 1988.

[101] D. H. Greene and D. E. Knuth. *Mathematics for the Analysis of Algorithms*, volume 1 of *Progress in Computer Science and Applied Logic (PCS)*. Birkhäuser Boston, 1990.

[102] J. Gross. *Mass Spectrometry: A textbook*. Springer, Berlin, 2004.

[103] K. Grützmann, S. Böcker and S. Schuster. Combinatorics of aliphatic amino acids. *Naturwissenschaften*, 98(1):79–86, 2011.

[104] M. Guilhaus. Principles and instrumentation in time-of-flight mass spectrometry. *J. Mass Spectrom.*, 30:1519–1532, 1995.

[105] S. Guillemot and F. Sikora. Finding and counting vertex-colored subtrees. In *Proc. of Symposium on Mathematical Foundations of Computer Science (MFCS 2010)*, volume 6281 of *Lect. Notes Comput. Sc.*, pages 405–416. Springer, 2010.

[106] C. Hamm, W. Wilson and D. Harvan. Peptide sequencing program. *Comput. Appl. Biosci.*, 2:115–118, 1986.

[107] F. Harary, R. W. Robinson and A. J. Schwenk. Twenty-step algorithm for determining the asymptotic number of trees of various species. *J. Austral. Math. Soc.*, 20(Series A): 483–503, 1975.

[108] M. Havilio, Y. Haddad and Z. Smilansky. Intensity-based statistical scorer for tandem mass spectrometry. *Anal. Chem.*, 75:435–444, 2003.

[109] M. Heinonen, A. Rantanen, T. Mielikäinen, J. Kokkonen, J. Kiuru, R. A. Ketola and J. Rousu. FiD: a software for ab initio structural identification of product ions from tandem mass spectrometric data. *Rapid Commun. Mass Spectrom.*, 22(19):3043–3052, 2008.

[110] D. W. Hill, T. M. Kertesz, D. Fontaine, R. Friedman and D. F. Grant. Mass spectral metabonomics beyond elemental formula: Chemical database querying by matching experimental with computational fragmentation spectra. *Anal. Chem.*, 80(14):5574–5582, 2008.

[111] W. M. Hines, A. M. Falick, A. L. Burlingame and B. W. Gibson. Pattern-based algorithm for peptide sequencing from tandem high energy collision-induced dissociation mass spectra. *J. Am. Soc. Mass Spectrom.*, 3(4):326 – 336, 1992.

[112] C. A. R. Hoare. FIND (algorithm 65). *Communications of the ACM*, 4:321–322, 1961.

[113] D. H. Horn, R. A. Zubarev and F. W. McLafferty. Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *J. Am. Soc. Mass Spectr.*, 11:320–332, 2000.

[114] C. S. Hsu. Diophantine approach to isotopic abundance calculations. *Anal. Chem.*, 56(8): 1356–1361, 1984.

[115] Q. Hu, R. J. Noll, H. Li, A. Makarov, M. Hardman and R. G. Cooks. The Orbitrap: a new mass spectrometer. *J. Mass Spectrom.*, 40(4):430–443, 2005.

[116] R. Hussong and A. Hildebrandt. Signal processing in proteomics. *Methods Mol. Biol.*, 604: 145–161, 2010.

[117] N. Jaitly, M. E. Monroe, V. A. Petyuk, T. R. W. Clauss, J. N. Adkins and R. D. Smith. Robust algorithm for alignment of liquid chromatography-mass spectrometry analyses in an accurate mass and time tag data analysis pipeline. *Anal. Chem.*, 78(21):7397–7409, 2006.

[118] N. Jeffries. Algorithms for alignment of mass spectrometry proteomic data. *Bioinformatics*, 21(14):3066–3073, 2005.

[119] R. S. Johnson and J. A. Taylor. Searching sequence databases via de novo peptide sequencing by tandem mass spectrometry. *Methods Mol. Biol.*, 146:41–61, 2000.

[120] R. S. Johnson and J. A. Taylor. Searching sequence databases via de novo peptide sequencing by tandem mass spectrometry. *Mol. Biotechnol.*, 22(3):301–315, 2002.

[121] P. Jones, R. G. Côté, L. Martens, A. F. Quinn, C. F. Taylor, W. Derache, H. Hermjakob, and R. Apweiler. PRIDE: a public repository of protein and peptide identifications for the proteomics community. *Nucleic Acids Res.*, 34(Database-Issue):659–663, 2006.

[122] H. J. Joshi, M. J. Harrison, B. L. Schulz, C. A. Cooper, N. H. Packer and N. G. Karlsson. Development of a mass fingerprinting tool for automated interpretation of oligosaccharide fragmentation data. *Proteomics*, 4(6):1650–1664, 2004.

[123] L. Käll, J. D. Canterbury, J. Weston, W. S. Noble and M. J. MacCoss. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods*, 4(11): 923–925, 2007.

[124] M. Kanehisa, S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, 34:D354–D357, 2006.

[125] R. Kannan. Lattice translates of a polytope and the Frobenius problem. *Combinatorica*, 12:161–177, 1991.

[126] E. A. Kapp, F. Schütz, L. M. Connolly, J. A. Chakel, J. E. Meza, C. A. Miller, D. Fenyo, J. K. Eng, J. N. Adkins, G. S. Omenn, and R. J. Simpson. An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: Sensitivity and specificity analysis. *Proteomics*, 5:3475–3490, 2005.

[127] M. Karas and F. Hillenkamp. Laser desorption ionization of proteins with molecular masses exceeding 10,000 Daltons. *Anal. Chem.*, 60:2299–2301, 1988.

[128] A. Keller, A. I. Nesvizhskii, E. Kolker and R. Aebersold. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.*, 74(20):5383–5392, 2002.

[129] A. Keller, J. Eng, N. Zhang, X.-J. Li and R. Aebersold. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol. Syst. Biol.*, 1:2005.0017, 2005.

[130] E. Kendrick. A mass scale based on CH2 = 14.0000 for high resolution mass spectrometry of organic compounds. *Anal. Chem.*, 35(13):2146–2154, 1963.

[131] A. Kerber, R. Laue and D. Moser. Ein Strukturgenerator für molekulare Graphen. *Anal. Chim. Acta*, 235:221 – 228, 1990.

[132] A. Kerber, R. Laue, M. Meringer and C. Rücker. Molecules in silico: The generation of structural formulae and its applications. *J. Comput. Chem. Japan*, 3(3):85–96, 2004.

[133] S. Kim, N. Gupta and P. A. Pevzner. Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *J. Proteome Res.*, 7(8):3354–3363, 2008.

[134] S. Kim, N. Bandeira and P. A. Pevzner. Spectral profiles, a novel representation of tandem mass spectra and their applications for de novo peptide sequencing and identification. *Mol. Cell. Proteomics*, 8(6):1391–1400, 2009.

[135] S. Kim, N. Gupta, N. Bandeira and P. A. Pevzner. Spectral dictionaries: Integrating de novo peptide sequencing with database search of tandem mass spectra. *Mol. Cell. Proteomics*, 8(1):53–69, 2009.

[136] T. Kind and O. Fiehn. Metabolomic database annotations via query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinformatics*, 7(1): 234, 2006.

[137] T. Kind and O. Fiehn. Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics*, 8:105, 2007.

[138] H. Kubinyi. Calculation of isotope distributions in mass spectrometry: A trivial solution for a non-trivial problem. *Anal. Chim. Acta*, 247:107–119, 1991.

[139] K.-S. Kwok, R. Venkataraghavan and F. W. McLafferty. Computer-aided interpretation of mass spectra. III. Self-training interpretive and retrieval system. *J. Am. Chem. Soc.*, 95 (13):4185–4194, 1973.

[140] V. Lacroix, C. G. Fernandes, and M.-F. Sagot. Motif search in graphs: Application to metabolic networks. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 3(4):360–368, 2006.

[141] A. J. Lapadula, P. J. Hatcher, A. J. Hanneman, D. J. Ashline, H. Zhang and V. N. Reinhold. Congruent strategies for carbohydrate sequencing. 3. OSCAR: an algorithm for assigning oligosaccharide topology from MS^n data. *Anal. Chem.*, 77(19):6271–6279, 2005.

[142] R. L. Last, A. D. Jones and Y. Shachar-Hill. Towards the plant metabolome and beyond. *Nat. Rev. Mol. Cell Biol.*, 8:167–174, 2007.

[143] A. Lavanchy, T. Varkony, D. H. Smith, N. A. B. Gray, W. C. White, R. E. Carhart, B. G. Buchanan, and C. Djerassi. Rule-based mass spectrum prediction and ranking: Applications to structure elucidation of novel marine sterols. *Org. Mass Spectrom.*, 15(7): 355–366, 1980.

[144] J. Lederberg. Topological mapping of organic molecules. *Proc. Natl. Acad. Sci. U. S. A.*, 53 (1):134–139, 1965.

[145] J. Lederberg. How DENDRAL was conceived and born. In *ACM Conference on the History of Medical Informatics, History of Medical Informatics archive*, pages 5–19, 1987. Available from `http://doi.acm.org/10.1145/41526.41528`.

[146] T. A. Lee. *A Beginner's Guide to Mass Spectral Interpretation*. Wiley, 1998.

[147] M. Lefmann, C. Honisch, S. Boecker, N. Storm, F. von Wintzingerode, C. Schloetelburg, A. Moter, D. van den Boom, and U. B. Goebel. A novel mass spectrometry based tool for genotypic identification of mycobacteria. *J. Clin. Microbiol.*, 42(1):339–346, 2004.

[148] G. Li and F. Ruskey. The advantages of forward thinking in generating rooted and free trees. In *Proc. of ACM-SIAM Symposium on Discrete Algorithms (SODA 1999)*, pages 939–940, Philadelphia, PA, USA, 1999. Society for Industrial and Applied Mathematics.

[149] G. Liu, J. Zhang, B. Larsen, C. Stark, A. Breitkreutz, Z.-Y. Lin, B.-J. Breitkreutz, Y. Ding, K. Colwill, A. Pasculescu, T. Pawson, J. L. Wrana, A. I. Nesvizhskii, B. Raught, M. Tyers, and A.-C. Gingras. ProHits: integrated software for mass spectrometry-based interaction proteomics. *Nat. Biotechnol.*, 28(10):1015–1017, 2010.

[150] K. K. Lohmann and C.-W. von der Lieth. GlycoFragment and GlycoSearchMS: web tools to support the interpretation of mass spectra of complex carbohydrates. *Nucleic Acids Res.*, 32(Web Server issue):W261–W266, 2004.

[151] B. Lu and T. Chen. A suffix tree approach to the interpretation of tandem mass spectra: Applications to peptides of non-specific digestion and post-translational modifications. *Bioinformatics*, 19(Suppl 2):ii113–ii121, 2003. Proc. of *European Conference on Computational Biology* (ECCB 2003).

[152] A. Luedemann, K. Strassburg, A. Erban and J. Kopka. TagFinder for the quantitative analysis of gas chromatography–mass spectrometry (GC-MS)-based metabolite profiling experiments. *Bioinformatics*, 24(5):732–737, 2008.

[153] G. S. Lueker. Two NP-complete problems in nonnegative integer programming. Technical Report TR-178, Department of Electrical Engineering, Princeton University, 1975.

[154] Y.-R. Luo. *Handbook of Bond Dissociation Energies in Organic Compounds*. CRC Press, Boca Raton, 2003.

[155] B. Ma and G. Lajoie. Improving the de novo sequencing accuracy by combining two independent scoring functions in peaks software. Poster at the ASMS Conference on Mass Spectrometry and Allied Topics, 2005.

[156] B. Ma, K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby and G. Lajoie. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.*, 17(20):2337–2342, 2003.

[157] B. Ma, K. Zhang and C. Liang. An effective algorithm for peptide de novo sequencing from MS/MS spectra. *J. Comput. Syst. Sci.*, 70:418–430, 2005.

[158] K. Maass, R. Ranzinger, H. Geyer, C.-W. von der Lieth and R. Geyer. "Glyco-peakfinder" – de novo composition analysis of glycoconjugates. *Proteomics*, 7(24):4435–4444, 2007.

[159] P. Mallick, M. Schirle, S. S. Chen, M. R. Flory, H. Lee, D. Martin, J. Ranish, B. Raught, R. Schmitt, T. Werner, B. Kuster, and R. Aebersold. Computational prediction of proteotypic peptides for quantitative proteomics. *Nat. Biotechnol.*, 25(1):125–131, 2007.

[160] M. Mann and M. Wilm. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.*, 66(24):4390–4399, 1994.

[161] S. Martello and P. Toth. An exact algorithm for large unbounded knapsack problems. *Oper. Res. Lett.*, 9(1):15–20, 1990.

[162] S. Martello and P. Toth. *Knapsack Problems: Algorithms and Computer Implementations*. John Wiley & Sons, Chichester, 1990.

[163] R. Matthiesen, J. Bunkenborg, A. Stensballe, O. N. Jensen, K. G. Welinder and G. Bauw. Database-independent, database-dependent, and extended interpretation of peptide mass spectra in VEMS V2.0. *Proteomics*, 4(9):2583–2593, 2004.

[164] R. Matthiesen, M. B. Trelle, P. Hojrup, J. Bunkenborg and O. N. Jensen. VEMS 3.0: algorithms and computational tools for tandem mass spectrometry based identification of post-translational modifications in proteins. *J. Proteome Res.*, 4(6):2338–2347, 2005.

[165] L. McHugh and J. W. Arthur. Computational methods for protein identification from mass spectrometry data. *PLoS Comput. Biol.*, 4(2):e12, 2008.

[166] P. E. Miller and M. B. Denton. The quadrupole mass filter: Basic operating concepts. *J. Chem. Educ.*, 63:617–622, 1986.

[167] L. Mo, D. Dutta, Y. Wan and T. Chen. MSNovo: a dynamic programming algorithm for de novo peptide sequencing via tandem mass spectrometry. *Anal. Chem.*, 79(13):4870–4878, 2007.

[168] E. Mostacci, C. Truntzer, H. Cardot and P. Ducoroy. Multivariate denoising methods combining wavelets and principal component analysis for mass spectrometry data. *Proteomics*, 10(14):2564–2572, 2010.

[169] I. K. Mun and F. W. Mclafferty. Computer methods of molecular structure elucidation from unknown mass spectra. In *Supercomputers in Chemistry*, ACS Symposium Series, chapter 9, pages 117–124. American Chemical Society, 1981.

[170] S. Na, J. Jeong, H. Park, K.-J. Lee and E. Paek. Unrestrictive identification of multiple post-translational modifications from tandem mass spectrometry using an error-tolerant algorithm based on an extended sequence tag approach. *Mol. Cell. Proteomics*, 7(12): 2452–2463, 2008.

[171] S. Neumann and S. Böcker. Computational mass spectrometry for metabolomics – a review. *Anal. Bioanal. Chem.*, 398(7):2779–2788, 2010.

[172] N. Nguyen, H. Huang, S. Oraintara and A. Vo. Mass spectrometry data processing using zero-crossing lines in multi-scale of Gaussian derivative wavelet. *Bioinformatics*, 26(18): i659–i665, 2010.

[173] R. Niedermeier. *Invitation to Fixed-Parameter Algorithms*. Oxford University Press, 2006.

[174] J. A. November. *Digitizing life: the introduction of computers to biology and medicine*. PhD thesis, Princeton University, Princeton, USA, 2006.

[175] H. Oberacher, M. Pavlic, K. Libiseller, B. Schubert, M. Sulyok, R. Schuhmacher, E. Csaszar, and H. C. Köfeler. On the inter-instrument and inter-laboratory transferability of a tandem mass spectral reference library: 1. results of an austrian multicenter study. *J. Mass Spectrom.*, 44(4):485–493, 2009.

[176] H. Oberacher, M. Pavlic, K. Libiseller, B. Schubert, M. Sulyok, R. Schuhmacher, E. Csaszar, and H. C. Köfeler. On the inter-instrument and the inter-laboratory transferability of a tandem mass spectral reference library: 2. optimization and characterization of the search algorithm. *J. Mass Spectrom.*, 44(4):494–502, 2009.

[177] S. Orchard, L. Montechi-Palazzi, E. W. Deutsch, P.-A. Binz, A. R. Jones, N. Paton, A. Pizarro, D. M. Creasy, J. Wojcik, and H. Hermjakob. Five years of progress in the standardization of proteomics data: 4th annual spring workshop of the HUPO-proteomics standards initiative. *Proteomics*, 7:3436–3440, 2007.

[178] R. Otter. The number of trees. *The Annals of Mathematics*, 49(3):583–599, 1948.

[179] K. G. Owens. Application of correlation analysis techniques to mass spectral data. *Appl. Spectrosc. Rev.*, 27(1):1–49, 1992.

[180] N. H. Packer, C.-W. von der Lieth, K. F. Aoki-Kinoshita, C. B. Lebrilla, J. C. Paulson, R. Raman, P. Rudd, R. Sasisekharan, N. Taniguchi, and W. S. York. Frontiers in glycomics: bioinformatics and biomarkers in disease. An NIH white paper prepared from discussions by the focus groups at a workshop on the NIH campus, Bethesda MD (September 11-13, 2006). *Proteomics*, 8(1):8–20, 2008.

[181] G. Palmisano, D. Antonacci and M. R. Larsen. Glycoproteomic profile in wine: a 'sweet' molecular renaissance. *J. Proteome Res.*, 9(12):6148–6159, 2010.

[182] D. J. Pappin, P. Hojrup and A. Bleasby. Rapid identification of proteins by peptide-mass fingerprinting. *Curr. Biol.*, 3(6):327–332, 1993.

[183] C. Y. Park, A. A. Klammer, L. Käll, M. J. MacCoss and W. S. Noble. Rapid and accurate peptide identification from tandem mass spectra. *J. Proteome Res.*, 7(7):3022–3027, 2008.

[184] W. E. Parkins. The uranium bomb, the calutron, and the space-charge problem. *Physics Today*, 58(5):45–51, 2005.

[185] V. Pellegrin. Molecular formulas of organic compounds: the nitrogen rule and degree of unsaturation. *J. Chem. Educ.*, 60(8):626–633, 1983.

[186] D. N. Perkins, D. J. Pappin, D. M. Creasy and J. S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18):3551–3567, 1999.

[187] R. H. Perry, R. G. Cooks and R. J. Noll. Orbitrap mass spectrometry: instrumentation, ion motion and applications. *Mass Spectrom. Rev.*, 27(6):661–699, 2008.

[188] G. Pólya. Kombinatorische Anzahlbestimmungen für Gruppen, Graphen und chemische Verbindungen. *Acta Mathematica*, 68(1):145–254, 1937.

[189] S. C. Pomerantz, J. A. Kowalak and J. A. McCloskey. Determination of oligonucleotide composition from mass spectromectrically measured molecular weight. *J. Am. Soc. Mass Spectrom.*, 4:204–209, 1993.

[190] R. Raman, S. Raguram, G. Venkataraman, J. C. Paulson and R. Sasisekharan. Glycomics: an integrated systems approach to structure-function relationships of glycans. *Nat. Methods*, 2(11):817–824, 2005.

[191] J. L. Ramírez-Alfonsín. *The Diophantine Frobenius Problem*. Oxford University Press, 2005.

[192] J. L. Ramírez-Alfonsín. Complexity of the Frobenius problem. *Combinatorica*, 16(1):143–147, 1996.

[193] I. Rauf, F. Rasche and S. Böcker. Computing maximum colorful subtrees in practice. Manuscript. **[TODO: REMOVE OR UPDATE]**, 2011.

[194] A. L. Rockwood and P. Haimi. Efficient calculation of accurate masses of isotopic peaks. *J. Am. Soc. Mass Spectrom.*, 17(3):415–419, 2006.

[195] A. L. Rockwood, M. M. Kushnir and G. J. Nelson. Dissociation of individual isotopic peaks: Predicting isotopic distributions of product ions in MS^n. *J. Am. Soc. Mass Spectr.*, 14:311–322, 2003.

[196] A. L. Rockwood, J. R. Van Orman and D. V. Dearden. Isotopic compositions and accurate masses of single isotopic peaks. *J. Am. Soc. Mass Spectr.*, 15:12–21, 2004.

[197] P. Roepstorff and J. Fohlman. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed. Mass Spectrom.*, 11(11):601, 1984.

[198] S. Rogers, R. A. Scheltema, M. Girolami and R. Breitling. Probabilistic assignment of formulas to mass peaks in metabolomics experiments. *Bioinformatics*, 25(4):512–518, 2009.

[199] R. G. Sadygov and J. R. Yates III. A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. *Anal. Chem.*, 75(15):3792–3798, 2003.

[200] R. G. Sadygov, D. Cociorva and J. R. Yates III. Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nat. Methods*, 1(3): 195–202, 2004.

[201] T. Sakurai, T. Matsuo, H. Matsuda and I. Katakuse. PAAS 3: A computer program to determine probable sequence of peptides from mass spectrometric data. *Biomed. Mass Spectrom.*, 11(8):396–399, 1984.

[202] A. Salomaa. Counting (scattered) subwords. *B. Euro. Assoc. Theo. Comp. Sci.*, 81:165–179, 2003.

[203] F. Sanger, S. Nicklen and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.*, 74(12):5463–5467, 1977.

[204] M. M. Savitski, M. L. Nielsen, F. Kjeldsen and R. A. Zubarev. Proteomics-grade de novo sequencing approach. *J. Proteome Res.*, 4:2348–2354, 2005.

[205] K. Scheubert, F. Hufsky, F. Rasche and S. Böcker. Computing fragmentation trees from metabolite multiple mass spectrometry data. In *Proc. of Research in Computational Molecular Biology (RECOMB 2011)*, volume 6577 of *Lect. Notes Comput. Sc.*, pages 377–391. Springer, 2011.

[206] J. Seidler, N. Zinn, M. E. Boehm and W. D. Lehmann. De novo sequencing of peptides by MS/MS. *Proteomics*, 10(4):634–649, 2010.

[207] J. Senior. Partitions and their representative graphs. *Am. J. Math.*, 73(3):663–689, 1951.

[208] B. Shan, B. Ma, K. Zhang and G. Lajoie. Complexities and algorithms for glycan sequencing using tandem mass spectrometry. *J. Bioinformatics and Computational Biology*, 6(1): 77–91, 2008.

[209] Q. Sheng, Y. Mechref, Y. Li, M. V. Novotny and H. Tang. A computational approach to characterizing bond linkages of glycan isomers using matrix-assisted laser desorption/ionization tandem time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.*, 22(22):3561–3569, 2008.

[210] I. V. Shilov, S. L. Seymour, A. A. Patel, A. Loboda, W. H. Tang, S. P. Keating, C. L. Hunter, L. M. Nuwaysir, and D. A. Schaeffer. The paragon algorithm, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra. *Mol. Cell. Proteomics*, 6(9):1638–1655, 2007.

[211] H. Shin, M. P. Sampat, J. M. Koomen and M. K. Markey. Wavelet-based adaptive denoising and baseline correction for MALDI TOF MS. *OMICS*, 14(3):283–295, 2010.

[212] F. Sikora. An (almost complete) state of the art around the graph motif problem. Technical report, Université Paris-Est, France, 2010. Available from `http://www-igm.univ-mlv.fr/~fsikora/pub/GraphMotif-Resume.pdf`.

[213] R. M. Silverstein, F. X. Webster and D. Kiemle. *Spectrometric Identification of Organic Compounds*. Wiley, 7th edition, 2005.

[214] G. Siuzdak. *The Expanding Role of Mass Spectrometry in Biotechnology*. MCC Press, second edition, 2006.

[215] D. H. Smith, N. A. Gray, J. G. Nourse and C. W. Crandell. The DENDRAL project: recent advances in computer-assisted structure elucidation. *Anal. Chim. Acta*, 133(4):471 – 497, 1981.

[216] R. K. Snider. Efficient calculation of exact mass isotopic distributions. *J. Am. Soc. Mass Spectrom.*, 18(8):1511–1515, 2007.

[217] H. M. Sobell. Actinomycin and DNA transcription. *Proc. Natl. Acad. Sci. U. S. A.*, 82(16): 5328–5331, 1985.

[218] H. Steen and M. Mann. The ABC's (and XYZ's) of peptide sequencing. *Nature Rev.*, 5: 699–711, 2004.

[219] M. T. Sykes and J. R. Williamson. Envelope: interactive software for modeling and fitting complex isotope distributions. *BMC Bioinformatics*, 9:446, 2008.

[220] J. J. Sylvester and W. J. Curran Sharp. Problem 7382. *Educational Times*, 37:26, 1884.

[221] D. L. Tabb, M. J. MacCoss, C. C. Wu, S. D. Anderson and J. R. Yates. Similarity among tandem mass spectra from proteomic experiments: detection, significance, and utility. *Anal. Chem.*, 75(10):2470–2477, 2003.

[222] H. Tang, Y. Mechref and M. V. Novotny. Automated interpretation of MS/MS spectra of oligosaccharides. *Bioinformatics*, 21 Suppl 1:i431–i439, 2005. Proc. of *Intelligent Systems for Molecular Biology* (ISMB 2005).

[223] S. Tanner, H. Shu, A. Frank, L.-C. Wang, E. Zandi, M. Mumby, P. A. Pevzner, and V. Bafna. Inspect: Identification of posttranslationally modified peptides from tandem mass spectra. *Anal. Chem.*, 77:4626–4639, 2005.

[224] J. A. Taylor and R. S. Johnson. Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry. *Anal. Chem.*, 73(11):2594–2604, 2001.

[225] J. A. Taylor and R. S. Johnson. Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.*, 11:1067–1075, 1997.

[226] J. van Lint and R. Wilson. *A Course in Combinatorics*. Cambridge University Press, 2001.

[227] A. Varki, R. D. Cummings, J. D. Esko, H. H. Freeze, P. Stanley, C. R. Bertozzi, G. W. Hart, and M. E. Etzler, editors. *Essentials of Glycobiology*. Cold Spring Harbor Laboratory Press, second edition, 2009. Freely available from `http://www.ncbi.nlm.nih.gov/books/NBK1908/`.

[228] R. Venkataraghavan, F. W. McLafferty and G. E. van Lear. Computer-aided interpretation of mass spectra. *Org. Mass Spectrom.*, 2(1):1–15, 1969.

[229] C.-W. von der Lieth, A. Bohne-Lang, K. K. Lohmann and M. Frank. Bioinformatics for glycomics: status, methods, requirements and perspectives. *Brief. Bioinform.*, 5(2):164–178, 2004.

[230] S. A. Waksman and H. B. Woodruff. Bacteriostatic and bacteriocidal substances produced by soil actinomycetes. *Proc. Soc. Exper. Biol.*, 45:609–614, 1940.

[231] M. S. Waterman and M. Vingron. Rapid and accurate estimates of statistical significance for sequence data base searches. *Proc. Natl. Acad. Sci. U. S. A.*, 91(11):4625–4628, 1994.

[232] J. T. Watson and O. D. Sparkman. *Introduction to Mass Spectrometry: Instrumentation, Applications, and Strategies for Data Interpretation*. Wiley, 2007.

[233] M. E. Wieser. Atomic weights of the elements 2005 (IUPAC technical report). *Pure Appl. Chem.*, 78(11):2051–2066, 2006.

[234] H. Wilf. *generatingfunctionology*. Academic Press, second edition, 1994. Freely available from `http://www.math.upenn.edu/~wilf/DownldGF.html`.

[235] S. Wolf, S. Schmidt, M. Müller-Hannemann and S. Neumann. In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics*, 11:148, 2010.

[236] W. E. Wolski, M. Lalowski, P. Jungblut and K. Reinert. Calibration of mass spectrometric peptide mass fingerprint data without specific external or internal calibrants. *BMC Bioinformatics*, 6:203, 2005.

[237] J. W. Wong, G. Cagney and H. M. Cartwright. SpecAlign—processing and alignment of mass spectra datasets. *Bioinformatics*, 21(9):2088–2090, 2005.

[238] L.-C. Wu, H.-H. Chen, J.-T. Horng, C. Lin, N. E. Huang, Y.-C. Cheng and K.-F. Cheng. A novel preprocessing method using Hilbert Huang transform for MALDI-TOF and SELDI-TOF mass spectrometry data. *PLoS One*, 5(8):e12493, 2010.

[239] Y. Wu, Y. Mechref, I. Klouckova, M. V. Novotny and H. Tang. A computational approach for the identification of site-specific protein glycosylations through ion-trap mass spectrometry. In *Proc. of RECOMB 2006 satellite workshop on Systems biology and computational proteomics*, volume 4532 of *Lect. Notes Comput. Sc.*, pages 96–107. Springer, 2007.

[240] C. Xu and B. Ma. Complexity and scoring function of MS/MS peptide de novo sequencing. In *Proc. of Computational Systems Bioinformatics Conference (CSB 2006)*, volume 4 of *Series on Advances in Bioinformatics and Computational Biology*, pages 361–369. Imperial College Press, 2006.

[241] J. Yates, P. Griffin, L. Hood and J. Zhou. Computer aided interpretation of low energy MS/MS mass spectra of peptides. In J. Villafranca, editor, *Techniques in Protein Chemistry II*, pages 477–485. Academic Press, San Diego, 1991.

[242] J. A. Yergey. A general approach to calculating isotopic distributions for mass spectrometry. *Int. J. Mass Spectrom. Ion Phys.*, 52(2–3):337–349, 1983.

[243] J. Zaia. Mass spectrometry of oligosaccharides. *Mass Spectrom. Rev.*, 23(3):161–227, 2004.

[244] J. Zhang, E. Gonzalez, T. Hestilow, W. Haskins and Y. Huang. Review of peak detection algorithms in liquid-chromatography-mass spectrometry. *Curr. Genomics*, 10(6):388–401, 2009.

[245] J. Zhang, D. Xu, W. Gao, G. Lin and S. He. Isotope pattern vector based tandem mass spectral data calibration for improved peptide and protein identification. *Rapid Commun. Mass Spectrom.*, 23(21):3448–3456, 2009.

[246] N. Zhang, R. Aebersold and B. Schwikowski. ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *Proteomics*, 2(10):1406–1412, 2002.

[247] W. Zhang and B. T. Chait. ProFound: an expert system for protein identification using mass spectrometric peptide mapping information. *Anal. Chem.*, 72(11):2482–2489, 2000.

[248] R. Zubarev and M. Mann. On the proper use of mass accuracy in proteomics. *Mol. Cell. Proteomics.*, 6(3):377–381, 2007.