Available online at www.sciencedirect.com

ScienceDirect

Journal of Discrete Algorithms 5 (2007) 714–728

JOURNAL OF DISCRETE ALGORITHMS

www.elsevier.com/locate/jda

Mass spectra alignments and their significance

Sebastian Böcker [a,*], Hans-Michael Kaltenbach [b,*]

[a] *Lehrstuhl für Bioinformatik, Friedrich-Schiller-Universität Jena, Ernst-Abbe-Platz 2, 07743 Jena, Germany*
[b] *AG Genominformatik, Technische Fakultät, Universität Bielefeld, PF 100 131, 33501 Bielefeld, Germany*

Received 5 December 2005; received in revised form 16 October 2006; accepted 6 November 2006

Available online 14 February 2007

Abstract

Mass Spectrometry has become one of the most popular analysis techniques in Genomics and Systems Biology. We investigate a general framework that allows the alignment (or matching) of any two mass spectra. In particular, we examine the alignment of a reference mass spectrum generated *in silico* from a database, with a measured sample mass spectrum. In this context, we assess the significance of alignment scores for character-specific cleavage experiments, such as tryptic digestion of amino acids. We present an efficient approach to estimate this significance, with runtime linear in the number of detected peaks. In this context, we investigate the probability that a random string over a weighted alphabet contains a substring of some given weight.

Keywords: Mass spectrometry; Fragment statistics; Global alignment; Protein identification

1. Introduction

Mass Spectrometry is one of the most popular analysis techniques in the emerging field of Systems Biology [1]: The analysis of peptide fingerprints and tandem mass spectra for protein identification and de novo sequencing is performed daily in thousands of laboratories around the world. The efficiency of this analysis technique is mainly due to its unique accuracy: Masses of sample molecules can be determined with an accuracy of parts of a neutron mass.

One central problem in the interpretation of mass spectrometry data is the matching of mass spectra: Usually, we are given some sample mass spectrum and a set of reference mass spectra (typically generated *in silico* from a sequence database), and we want to know which of the reference mass spectra fits best to the sample mass spectrum. To compute such peak matchings, we investigate a general framework that allows to *align* any two spectra and give examples on how to score such alignments.

As the other main contribution of this paper, we assess the quality of such an alignment: We develop a framework for efficiently computing p-values for restriction-type experiments, such as tryptic digestion of amino acids also known as peptide mass fingerprinting [2]. We also report preliminary results for tandem mass spectrometry data (MS/MS, see [3]). Here one often wants to determine the amino acid sequence without using a sequence database (peptide de novo sequencing); but then, we can use our approach of aligning mass spectra to accurately discriminate between candidate sequences generated by an de novo sequencing algorithm [4].

[*] Corresponding author.
E-mail addresses: boecker@minet.uni-jena.de (S. Böcker), michael@cebitec.uni-bielefeld.de (H.-M. Kaltenbach).

1570-8667/$ – see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.jda.2006.11.003

In particular, we are interested in the question whether a random string over a weighted alphabet contains a (certain) substring of given weight. This question has been frequently addressed using heuristics and approximations [5,6] in order to analyze mass spectrometry data. We present a surprisingly simple recurrence relation that allows exact and efficient computation of these fragment occurrence probabilities.

We believe that the two-step process of first aligning spectra, then assessing the significance of the alignment—which has proven useful in the context of string alignments—is also beneficial for the analysis of mass spectrometry data. First, the flexibility of scoring schemes allows to adjust to an application's peculiarities with a maximal degree of freedom. Second, certain scoring schemes emerge quite naturally in a statistical context. Third, alignment scores themselves are often quite useless for protein identification, because long proteins usually achieve better scores than short ones. Our approach allows an efficient estimation of the p-value of an alignment score taking into account protein lengths, and therefore combines the advantages of alignments and stochastic analysis. Note that MASCOT [5,7], the most popular program for peptide mass fingerprinting, approaches this problem by a heuristic estimation of such p-values.

Related work. The problem of aligning mass spectra is clearly related to the well-known Longest Common Subsequence Problem.

Similar approaches were used for physical map comparison [8] and aligning gel electrophoresis patterns [9,10], as well as matching tree ring data [11], but only [12] contains a comparable approach for aligning mass spectra, and uses edit distances with restricted gaps. We would like to stress that there is no correspondence between our approach, and "spectral alignments" introduced in [13].

The significance or probability of mass spectrum matchings has been considered using a two-step stochastic process [14], a hypothesis testing formulation [15], and using empirical statistics [16].

2. Definition of the model

Solely for readability, we limit our attention to ionization methods that predominantly produce single charged ions, such as MALDI [17]. This allows us to talk about the mass m of a molecule, instead of its mass-to-charge ratio m/z.

We can compute the mass of a biomolecule *in silico*, simply summing up the masses of its atoms. Mass spectrometry allows to estimate molecule masses with an exceptionally high accuracy such as 0.1 Dalton (Da), about one tenth the mass of a neutron. Still and all, measured masses usually differ from those theoretically predicted.

When comparing simulated and measured mass spectra, we have to take into account the *resolution* constraint of mass spectrometers: In theory, glutamine residues with sum formula $C_5H_8N_2O_2$ have a mass of 128.1315... Da, while lysine residues with sum formula $C_6H_{12}N_2O$ have a mass of 128.1752... Da, for natural isotopic distribution. If two molecules have almost identical masses, the corresponding peaks may overlap in the measured mass spectrum, and ultimately create a joint peak with mass somewhere in-between the two original masses. This effect is hard to predict *in silico*, because it is highly dependent on a multitude of parameters, such as mass spectrometer settings, and one usually predicts the reference spectrum ignoring this effect.

In the following, we further simplify matters by assuming that a *mass spectrum* is a set of peaks. Every peak has a *mass*, and eventually other attributes such as intensity, signal-to-noise ratio or area-under-curve.

Definition 1. A *mass spectrum* or *peak-list* of length n is a list $\mathcal{S} = \{p_1, \ldots, p_n\}$ of *peaks* $p_i \in \mathcal{M} \times \mathcal{A}$. Every such peak has a *mass* $m_i \in \mathcal{M} \subseteq \mathbb{R}$ and possibly other attributes $(a_{i,1}, \ldots, a_{i,k}) \in \mathcal{A}$, $k \geqslant 0$. A spectrum is sorted by mass, that is, $m_i < m_j$ whenever $i < j$, $1 \leqslant i, j \leqslant |\mathcal{S}|$.

Working with peak lists is a widely used simplification when analyzing mass spectra, and implies that an efficient peak detection algorithm has been applied to the raw data of a sample mass spectrum, differentiating between peaks and background "noise".

Example 2. The simplest representation of a peak is its mass. Then $\mathcal{M} = \mathbb{R}$ and $\mathcal{A} = \emptyset$. If we also like to consider the relative intensity of a peak (compared to the other peaks in the same spectrum), we could set $\mathcal{M} = \mathbb{R}$ and $\mathcal{A} = [0, 1]$. A peak p_i would then be a tuple $(m_i, a_{i,1})$ of mass m_i and intensity $a_{i,1}$.

Due to the imperfection of mass spectrometry and biochemistry, a sample mass spectrum may differ from an ideal mass spectrum and show *additional* and *missing* peaks.

Mass spectrometry measures the masses of sample molecules. For peptides as well as nucleotides, these molecules can be viewed as strings over the (amino acid or nucleic acid) alphabet, and the weight of a string is simply the sum of weights of its characters.

Definition 3. A *weighted alphabet* Σ is a finite alphabet together with a *character weight function* $\mu : \Sigma \to \mathbb{R}$. We can extend the domain of μ to strings $s = s_1 \dots s_n \in \Sigma^*$ by defining $\mu(s) := \sum_{i=1}^{n} \mu(s_i)$. We set $\mu_{\min} := \min_{\sigma \in \Sigma} \mu(\sigma)$ and $\mu_{\max} := \max_{\sigma \in \Sigma} \mu(\sigma)$ for the smallest and greatest character mass in Σ.

Depending on the experimental settings, there exists a maximal mass $m_{\max} \in \mathbb{R}$ such that no masses above $m_{\max}$ are present in any mass spectrum: for example, $m_{\max} \approx 3000$ for tryptic digestion experiments. Then, $\mathcal{M} := [0, m_{\max}]$ is the peak *mass range* of interest, and $l_{\max} := \lfloor m_{\max}/\mu_{\min} \rfloor$ is the maximal length of a fragment that we can detect.

Example 4. Consider the amino acid alphabet $\Sigma = \{\mathrm{A}, \mathrm{C}, \dots, \mathrm{W}, \mathrm{Y}\}$. For natural isotopic distribution, the masses (in Dalton, four digits accuracy) of some of the characters are:

σ	A (Ala)	C (Cys)	D (Asp)	E (Glu)	…	W (Trp)	Y (Tyr)
$\mu(\sigma)$	71.0371	103.0092	115.0269	129.0426	…	186.0793	163.0633
int. $\mu(\sigma)$	710	1030	1150	1290	…	1861	1631

We will sometimes require that all masses are natural numbers. To this end, we round the true masses to integers using some mass accuracy $\Delta \in \mathbb{R}$: Above we have denoted integer masses for $\Delta = 0.1$. In this *discrete* case, $\mathcal{M} := \{0, \dots, m_{\max}\}$ is the mass range of interest, for example $\mathcal{M} = \{0, 1, \dots, 30\,000\}$ for tryptic digestion and accuracy $\Delta = 0.1$.

In Section 5, we investigate mass spectra that come from biochemical experiments involving character-specific cleavage, such as tryptic digestion of amino acids. A mathematical formalism for such cleavage was introduced in [18]. We restate the definition.

Definition 5. Given a sample string $s \in \Sigma^*$ and a *cleavage character* $x \in \Sigma$, a substring $y \in (\Sigma - \{x\})^*$ is a *fragment* of s if xyx is a substring of xsx.

Such fragments correspond to *complete* cleavage of the string, but the methods presented below can be easily extended to take into account partial as well as incomplete cleavage.

Given a reference string, it is straightforward to compute all masses in the corresponding reference spectrum [5,18]. We do not go into the details here and refer the reader to the literature.

The utilized biochemistry sometimes leads to mass modifications of fragment masses, such as +18 Da for an additional H_2O group, and to terminal fragments (corresponding to beginning and end of the sample string) that usually differ in mass from non-terminal fragments. Also, biological cleavage reactions for proteins usually have more than one cleavage character and the cleavage reaction is suppressed in the presence of certain prohibition characters, following the cleavage character in the sequence. We will ignore all these modifications of the model for readability. An extensive treatment how to extend the model can be found in [19].

In Section 7, we confine our analysis to collision-induced dissociation by tandem mass spectrometry. There, we break the peptide string s into all prefixes and suffixes of s.

3. Spectrum alignment

Let $\mathcal{S} = \{p_1, \dots, p_n\}$, $\mathcal{S}' = \{p'_1, \dots, p'_{n'}\}$ be the two spectra of length n and n', respectively, that we want to match. We want to construct a peak matching, that is: a bijective map $\pi : \mathcal{S}_* \to \mathcal{S}'_*$ where $\mathcal{S}_* \subseteq \mathcal{S}$ and $\mathcal{S}'_* \subseteq \mathcal{S}'$ are the matched peaks, while all other peaks remain unmatched. We assume that the map π is a bijection, but we describe below how to deal with many-to-one matchings.

To find the optimal matching of $\mathcal{S}$ and $\mathcal{S}'$, we have to assign scores to any possible matching using a scoring function *score* for single peak matching:

$$\mathit{score}: \big(\mathcal{S} \cup \{\varepsilon\}\big) \times \big(\mathcal{S}' \cup \{\varepsilon'\}\big) \to \mathbb{R}$$

where ε and ε' denote special "gap" peaks and $\mathit{score}(\varepsilon, \varepsilon') := -\infty$. Different gap peaks are needed to allow the two spectra to have different additional peak attributes.

For $p_i \in \mathcal{S}$ and $p'_j \in \mathcal{S}'$, $\mathit{score}(p_i, p'_j)$ is the score of matching peaks p_i in $\mathcal{S}$ and p'_j in $\mathcal{S}'$; $\mathit{score}(p_i, \varepsilon')$ is the score of a *missing* peak p_i in $\mathcal{S}$ not present in $\mathcal{S}'$; and $\mathit{score}(\varepsilon, p'_j)$ is the score of an *additional* peak p'_j in $\mathcal{S}'$ not present in $\mathcal{S}$.

In the following, we do not make any assumptions regarding the peak scoring function *score*. It is clear that such a scoring function must be based on the peaks attributes, such as mass or intensity: for example, if m_i is the mass of peak $p_i \in \mathcal{S}$ and m'_j the mass of peak $p'_j \in \mathcal{S}'$, then $\mathit{score}(p_i, p'_j)$ should be the higher, the smaller the mass difference $|m_i - m'_j|$ is. The presented framework allows us to mimic *any* additive or multiplicative scoring scheme, such as that used by MASCOT [5] or log likelihood peak scoring [20]. We will discuss some details of useful scoring schemes in the next section.

Now, the score of the matching $\pi : \mathcal{S}_* \to \mathcal{S}'_*$ is the sum of scores of the peak matchings:

$$\mathit{score}(\pi) = \sum_{p_i \in \mathcal{S}_*} \mathit{score}\big(p_i, \pi(p_i)\big) + \sum_{p_i \in \mathcal{S} \setminus \mathcal{S}_*} \mathit{score}(p_i, \varepsilon') + \sum_{p'_j \in \mathcal{S}' \setminus \mathcal{S}'_*} \mathit{score}(\varepsilon, p'_j). \tag{1}$$

We are searching for a maximal score among all matchings.

Example 6. Using only peak masses for scoring, we define a *peak counting score* by setting

$$\mathit{score}(p_i, p'_j) = \begin{cases} 1, & \text{if } |m_i - m'_j| \leqslant \delta \\ 0, & \text{otherwise} \end{cases}$$

for all $p_i \in \mathcal{S}$ and $p'_j \in \mathcal{S}'$ having masses m_i and m'_j, respectively, and for some fixed mass difference $\delta \in \mathbb{R}$. Setting gap scores $\mathit{score}(p_i, \varepsilon') = \mathit{score}(\varepsilon, p'_j) = 0$, we simply count the number of peaks we can match with a mass difference of at most δ.

To exclude meaningless matchings, we only allow *non-intersecting* peak matchings. Peak matchings are non-intersecting, if the following conditions holds:

$$m_i < m_j \quad \text{if and only if} \quad m'_{i'} < m'_{j'}$$

for all $p_i, p_j \in \mathcal{S}_*$ mapping $p'_{i'} = \pi(p_i)$ and $p'_{j'} = \pi(p_j)$. Since we require masses to be ordered in a spectrum, this monotonicity condition of masses is equivalent to a monotonicity condition of peak indices as for $p_i, p_j, p'_{i'}, p'_{j'}$ as above, we have that "$i < j \iff i' < j'$". In this sense, the bijection π is strictly monotonic and is hence uniquely determined by the choice of subsets $\mathcal{S}_* \subseteq \mathcal{S}$ and $\mathcal{S}'_* \subseteq \mathcal{S}'$.

These considerations show that we are searching for an *alignment* between the two spectra $\mathcal{S}$ and $\mathcal{S}'$. Computing the optimal, i.e. highest scoring, alignment can be done efficiently using Dynamic Programming, and we define the well-known recurrence relation for the $(n+1) \times (n'+1)$ matrix E by

$$\begin{aligned}
&E[0,0] = 0 \\
&E[i+1,0] = E[i,0] + \mathit{score}(p_{i+1}, \varepsilon') \\
&E[0,j'+1] = E[0,j'] + \mathit{score}(\varepsilon, p'_{j'+1}) \\
&E[i+1,j'+1] = \max \left\{ \begin{array}{l} E[i,j'+1] + \mathit{score}(p_{i+1}, \varepsilon'), \\ E[i+1,j'] + \mathit{score}(\varepsilon, p'_{j'+1}), \\ E[i,j'] + \mathit{score}(p_{i+1}, p'_{j'+1}) \end{array} \right\}
\end{aligned} \tag{2}$$

using the familiar boundary conditions. Now, $\mathit{score}(\mathcal{S}, \mathcal{S}') = E[n,n']$ holds the score of an optimal alignment between $\mathcal{S}, \mathcal{S}'$, and we can find all such optimal alignments by backtracking through the matrix E.

The incorporation of additional attributes like peak intensities may be of particular importance when scoring missing and additional peaks: For missing peaks, we have transformed the raw data of the mass spectrum into a peak list discarding candidates whose intensity falls below a given threshold. Hence, slight changes of this threshold can dramatically change scores that do not take into account peak intensities. For additional peaks, similar arguments apply.

It should be understood that for reasonable peak scorings, we do not have to fill in the complete matrix E: We can expect that $score(p_i, p'_j)$ decreases as the mass difference $|p_i - p'_j|$ increases. In particular, $score(p_i, p'_j)$ will be very small for high mass differences, because there is no reason to match two peaks that are, say, 1000 Da apart. On the contrary, scores $score(p_i, \varepsilon')$ and $score(\varepsilon, p'_j)$ are mostly independent of peak masses. Let θ be a lower bound of $score(p_i, \varepsilon')$ and $score(\varepsilon, p'_j)$. From the above, we may assume that there exists some mass difference δ such that $score(p, p') \leqslant 2\theta$ for all peaks with $|p - p'| \geqslant \delta$. So, it suffices to fill in only those parts of the matrix E where $|p_i - p'_j|$ is not too large. The optimal alignment can then be calculated by "banded" dynamic programming in time $\mathrm{O}(|C| + |\mathcal{S}| + |\mathcal{S}'|)$ where $C := \{(i, j){:}\ |p_i - p'_j| \leqslant \delta\}$ is the set of potential matches: for every peak p_i there exist indices j_0, j_1 such that $|p_i - p'_j| \leqslant \delta$ if and only if $j \in \{j_0, \dots, j_1\}$. Going from i to $i + 1$ we only have to increase the pointers j_0, j_1.

Example 7. Given two spectra $\mathcal{S} := \{p_1, \dots, p_4\}$ and $\mathcal{S}' := \{p'_1, \dots, p'_5\}$, let $\{m_1, \dots, m_4\} = \{200, 510, 705, 850\}$ and $\{m'_1, \dots, m'_5\} = \{200, 300, 500, 515, 700\}$ be their peak masses. For $\delta = 10$ and the "peak counting score" introduced in Example 6, one can easily calculate $E[4, 5] = 3$, so an optimal alignment matches three peaks.

$E[i, j']$	ε'	200	300	500	515	700
ε	0	0	0	0	0	0
200	0	1	1	1	1	1
510	0	1	1	2	2	2
705	0	1	1	2	2	3
850	0	1	1	2	2	3

If we use the slightly more complex function

$$score(p_i, p'_j) := 2 - \frac{1}{5}|m_i - m'_j| \quad \text{and} \quad score(p_i, \varepsilon') = score(\varepsilon, p'_j) = -1$$

for all i, j then the matrix $E[i, j']$ is:

$E[i, j']$	ε'	200	300	500	515	700
ε	**0**	−1	−2	−3	−4	−5
200	−1	**2**	**1**	**0**	−1	−2
510	−2	1	0	1	**1**	0
705	−3	0	−1	0	0	**2**
850	−4	−1	−2	−1	−1	**1**

For readability, we print masses m_i and m'_j instead of indices i and j in these tables. We have grayed out those entries of $E[i, j']$ that need not to be calculated. So, an optimal alignment has score $E[4, 5] = 1$; we can achieve this score matching $m_1 = 200$ with $m'_1 = 200$, $m_2 = 510$ with $m'_4 = 515$, and $m_3 = 705$ with $m'_5 = 700$.

4. Many-to-one peak matchings and scoring functions

We now concentrate on matching a single sample mass spectrum to a multitude of reference spectra generated *in silico*.

So far, we have not elaborated on how to choose peak score $score(\cdot, \cdot)$. To this end, we define a *global peak scoring function* $\Psi : (\mathcal{M} \times \mathcal{A}) \times (\mathcal{M} \times \mathcal{A}') \to \mathbb{R}$ that maps a reference peak $p \in \mathcal{M} \times \mathcal{A}$ and a sample peak $p' \in \mathcal{M} \times \mathcal{A}'$ to a peak score $\Psi(p, p')$. This map is independent of an actual reference or sample spectrum, and even actual peaks. Note that we do allow different additional peak attributes in the two spectra, but require the same mass range. Now, we can define $score(p, p') := \Psi(p, p')$ for $p \in \mathcal{S}$, $p' \in \mathcal{S}'$. In Example 7 we have implicitly used the global peak scoring function $\Psi(p, p') := 2 - \frac{1}{5}|m - m'|$.

Example 8. Assume that sample peak masses are normally distributed around the ideal peak mass m. The variance of this distribution may also depend on m, since large mass errors appear more often in high mass regions, but here we concentrate on a constant variance $\bar{\sigma}^2$. If we want to positively score, say, 95% of all sample peaks (so, the mass difference must be smaller than approximately $2\bar{\sigma}$), we can define

$$\Psi(p, p') := 1 - \mathbb{P}(Z > -z \text{ and } Z < z) \tag{3}$$

where $z := |m - m'|/\bar{\sigma}$ and $Z \sim \mathcal{N}(0, 1)$. Then, $\Psi(p, p) = 1$ and $\Psi(p, p') \approx 0$ holds for $|m - m'| = 2\bar{\sigma}$.

Choosing a "good" peak scoring function highly depends on the underlying application, and surely is a problem of its own.

We also have to score additional peaks by $score(\varepsilon, \cdot)$ and missing peaks by $score(\cdot, \varepsilon')$. To this end, let $\Psi^{\text{add}} : \mathcal{M} \times \mathcal{A}' \to \mathbb{R}$ and $\Psi^{\text{miss}} : \mathcal{M} \times \mathcal{A} \to \mathbb{R}$ be two functions that score an additional peak $p' \in \mathcal{S}'$, or a missing peak $p \in \mathcal{S}$. These functions can be defined constant as in Example 7, but as mentioned before, it is also reasonable to take into account peak intensities as well as experience about experimental settings.

We introduce some notations that will be of use when calculating the significance of an alignment score. Given a fixed sample spectrum $\mathcal{S}'$, we concentrate on a single sample peak $p'_j \in \mathcal{S}'$ and abbreviate:

$$\Psi_j : \mathcal{M} \times \mathcal{A} \to \mathbb{R}, \quad \text{where } \Psi_j(p) := \Psi(p, p'_j) \text{ for } p \in \mathcal{M} \times \mathcal{A}.$$

Similarly, we write $\Psi_j^{\text{add}} := \Psi^{\text{add}}(p'_j)$ for additional and $\Psi_i^{\text{miss}} := \Psi^{\text{miss}}(p_i)$ for missing peaks.

To simplify computations, we postulate that every peak scoring function has finite and compact *support*: That is, $\Psi_j(p)$ is above a certain threshold if and only if the mass m of p is inside the interval $[m_1, m_2]$ for masses m_1, m_2. In the discrete case, the support $\{m_1, \dots, m_2\}$ of Ψ_j is denoted $\mathcal{U}_j$, and reference peaks with mass $m \notin \mathcal{U}_j$ will never be matched to sample peak p'_j. We further require that the support of two peaks $p'_j, p'_{j+1} \in \mathcal{S}'$ does not intersect, and we can achieve this by shrinking overlapping support.

Often, we want to match a single sample peak to one *or more* reference peaks. The simplest incorporation of such many-to-one peak matchings is as follows: We simply add scores of matching a sample peak p'_j to all reference peaks p_i with mass $m_i \in \mathcal{U}_j$, and if there is no such reference peak, we score peak p'_j by Ψ_j^{add}. Now,

$$score_{\text{m2o}}(\mathcal{S}, \mathcal{S}') := \sum_{p'_j \in \mathcal{S}'} \sum_{p_i \in \mathcal{S},\, m_i \in \mathcal{U}_j} \Psi_j(p_i) + \sum_{p'_j \text{ additional}} \Psi_j^{\text{add}} + \sum_{p_i \text{ missing}} \Psi^{\text{miss}}(p_i) \tag{4}$$

where "p'_j additional" runs over those $p'_j \in \mathcal{S}'$ where there is no $p_i \in \mathcal{S}$ with $m_i \in \mathcal{U}_j$; and "p_i missing" runs over those $p_i \in \mathcal{S}$ where there is no $p'_j \in \mathcal{S}'$ with $m_i \in \mathcal{U}_j$. We can compute $score_{\text{m2o}}$ in time $\mathrm{O}(|\mathcal{S}| \cdot |\mathcal{S}'|)$, or $\mathrm{O}(|C| + |\mathcal{S}| + |\mathcal{S}'|)$ where C is again the set of potential matches.

For a particular reference spectrum $\mathcal{S}$, it is useful to take into account interferences if additional peak attributes such as intensity are known: Peak intensities are mostly additive, and a sample peak that is matched to two or more reference peaks should show an intensity that is the sum of intensities of the reference peaks. We can modify the spectrum alignment of Section 3 to take into account multiple matches, by trying to align a single sample peak to more than only the last reference peaks in the dynamic programming recurrence (2). Such *merging alignment* can be computed in time $\mathrm{O}(|\mathcal{S}| \cdot |\mathcal{S}'| \cdot k)$ where k denotes the maximal number of reference peaks with masses that fall into the support of any single sample peak. Omitting the details we just note that computations are usually faster than this worst-case runtime suggests.

5. Character-specific cleavage of random strings

We now concentrate on the case that the measured mass spectra come from biochemical experiments involving character-specific cleavage, such as tryptic digestion of amino acids or RNAse digestion of nucleotides, as defined in Definition 5.

For computing the distributions of alignment scores in order to assess the significance of identifications in Section 6, we need to compute the contribution of a peak p with mass m to an overall alignment score in the above setting.

That is, we need to compute the *occurrence probability* that at least one fragment of mass m occurs in a random string $s \in \Sigma^L$ of some given length L. We assume the characters of s to be drawn independently with uniform probabilities $1/|\Sigma|$. Generalizations of this model to other distributions and cleavage reactions can be found in [19].

Formally, let $\Sigma_x := \Sigma - \{x\}$ be the alphabet without cleavage character. For $m \in \mathbb{N}$ and $L \in \mathbb{N}$ let $S[L,m]$ be the set of strings of length L that have at least one fragment of mass m:

$$S[L,m] := \bigl\{ s \in \Sigma^L \colon s \text{ contains fragment } y \in \Sigma_x^* \text{ with } \mu(y) = m \bigr\}. \tag{5}$$

We want to compute the *occurrence probability* $p[L,m] = \mathbb{P}(s \in S[L,m])$ of a fragment of mass m in a random string $s \in \Sigma^L$. Again, let $m_{\max}$ be the largest mass to consider, $l_{\max}$ the length of the longest possible fragment and assume all masses to be integers.

For later use, we first compute the number $d[m]$ of fragments $y \in \Sigma_x^*$ having mass $\mu(y) = m$. It is computer science folklore that we can compute this number using the simple recurrence relation $d[0] := 1$ and

$$d[m] = \sum_{\sigma \in \Sigma,\, \mu(\sigma) \leqslant m} d\bigl[m - \mu(\sigma)\bigr] \quad \text{for } m \geqslant 0. \tag{6}$$

Computing $d[\cdot]$ takes $\mathrm{O}(|\Sigma| \cdot m_{\max})$ time, and storing it requires $\mathrm{O}(m_{\max})$ space.

Now, for a length l and a mass $m \in \mathbb{N}$, let $c[l,m]$ denote the number of strings $y \in \Sigma_x^l$ such that $\mu(y) = m$. We can compute $c[\cdot,\cdot]$ by initializing $c[0,0] := 1$, $c[0,m] := 0$ for all $m > 0$, and the recurrence relation

$$c[l,m] = \sum_{\sigma \in \Sigma_x,\, \mu(\sigma) \leqslant m} c\bigl[l-1, m - \mu(\sigma)\bigr]$$

for $l \geqslant 1$ and $m \geqslant 0$. Note that $c[l,m] = 0$ for $l < m/\mu_{\max}$ as well as for $l > m/\mu_{\min}$. Computing $c[\cdot,\cdot]$ takes $\mathrm{O}(|\Sigma| \cdot l_{\max} \cdot m_{\max})$ time, and storing it requires $\mathrm{O}(l_{\max} \cdot m_{\max})$ space.

Combining $c[l,m]$ with our random string model gives us the *fragment probability* $f[l,m]$ that a string $s \in \Sigma^*$ has a first fragment $y = s_{1:l}$ of mass m and length l.

Lemma 9. *The fragment probability $f[l,m]$ is given by*

$$f[l,m] := \mathbb{P}\bigl(s_{1:l} \in \Sigma_x^l,\, \mu(s_{1:l}) = m,\, s_{l+1} = x\bigr) = \frac{c[l,m]}{(|\Sigma|-1)^l} \cdot \frac{1}{|\Sigma|} \cdot \left(1 - \frac{1}{|\Sigma|}\right)^l$$

for $l < |s|$ and $f[|s|,m] = \frac{c[|s|,m]}{(|\Sigma|-1)^{|s|}} \cdot (1 - \frac{1}{|\Sigma|})^{|s|}$. The probability of the complementary event to have a fragment of length l not having mass m is

$$\bar{f}[l,m] = \left(1 - \frac{c[l,m]}{(|\Sigma|-1)^l}\right) \cdot \left(1 - \frac{1}{|\Sigma|}\right)^l \cdot \frac{1}{|\Sigma|},$$

and $\bar{f}[0,m] = 1/|\Sigma|$ in particular.

Proof. Recall that the cleavage character is not part of the fragment. Thus, the probability to have a first fragment of length l is the probability to see l non-cleavage characters directly followed by the cleavage character x. This probability is given from the geometric distribution by $(1 - 1/|\Sigma|)^l \cdot 1/|\Sigma|$. There are Σ_x^l strings of length l having no cleavage character. Among them, $c[l,m]$ have required mass m.

Clearly, the first fragment can only be as long as the string itself. In this case, the first fragment is identical with the string and no cleavage character is needed to end the fragment.

Fragments of length 0 have to be treated if a string starts with a cleavage character. □

Using the fragment probabilities, we can finally compute the occurrence probabilities by looking at the complementary event $\bar{p}[L,m] := 1 - p[L,m]$ to have no fragment of mass m in a random string of length L.

Theorem 10. *For a random string $s \in \Sigma^L$ and a fixed mass m, we can compute $\bar{p}[L,m]$ using the initial values $\bar{p}[0,m] = 1$, and the recurrence relation*

$$\bar{p}[L,m] = \bar{f}[L,m] + \sum_{l=1}^{L} \bar{p}[L-l,m] \cdot \bar{f}[l-1,m]. \tag{7}$$

Proof. Consider the empty string: Clearly, it does not have a fragment of mass m and thus the initial condition holds.

Let us denote the first occurrence of a cleavage character x in string s of length $L > 0$ by $t(x)$ and set $t(x) = L + 1$ if s does not contain a cleavage character. Then, the prefix $s_{1:t(x)-1}$ is the first fragment of s. Because of the independence of characters, the masses of fragments in the remaining suffix $s_{t(x)+1:L}$ are independent of the mass of the first fragment, given $t(x)$. The probability that s has no fragment of mass m is thus the product of the probability that its first fragment does not have mass m and the remaining suffix contains no fragment of this mass:

$$\bar{p}[L,m] = \bar{f}\bigl[t(x)-1,m\bigr] \cdot \bar{p}\bigl[L-t(x),m\bigr].$$

Summing over all possible values $1 \ldots L$ of $t(x)$ and explicitly taking care of the special case $t(x) = L + 1$ with the $\bar{f}[L,m]$ term gives the stated result. □

Eq. (7) is essentially a convolution over string-lengths to cover all possible lengths of the first fragment and the remaining suffix that sum up to length L. Moreover, the equation does not count strings twice; the length of the first fragment partitions the set of strings of length L into non-overlapping subsets.

Let us briefly investigate the particular case $L = 1$ in some more detail: The recurrence then reduces to $\bar{p}[1,m] = \bar{f}[1,m] + \bar{f}[0,m]$. The first term is the probability that the first (and only) fragment has length 1, i.e. is the string itself. The second term corresponds to $s = x$, i.e. the string is a cleavage character. This event has probability $1/|\Sigma|(= \bar{f}[0,m])$ independent of m, and the first fragment as well as the suffix are empty, i.e. both do not have mass m with probability 1.

A naive implementation of the recurrence given in Theorem 10 would require $\mathrm{O}(L_{\max}^2 \cdot m_{\max})$ time. It is however possible do exploit some dependencies among successive computations.

Lemma 11. *The occurrence probabilities can be computed in time $\mathrm{O}(L_{\max} \cdot l_{\max} \cdot m_{\max})$ using the recurrence equation of Theorem 10.*

Proof. Recall that $c[l,m] = 0$ for $l > l_{\max}$. In this case, $\bar{f}[l,m] = (1 - 1/|\Sigma|)^l \cdot 1/|\Sigma|$ and thus, $\bar{f}[l,m]$ is independent of m. Then, $\bar{f}[l,m] = (1 - 1/|\Sigma|) \cdot \bar{f}[l-1,m]$. Now consider the case that $\bar{p}[L,m]$ has been computed up to some $L > l_{\max}$. The next entry would then be $\bar{p}[L+1,m] = \bar{f}[L,m] + \sum_{l=1}^{L+1} \bar{p}[L+1-l,m] \cdot \bar{f}[l-1,m]$. We can split the sum in a part to $l_{\max}$ and the rest:

$$\bar{p}[L+1,m] = \bar{f}[L,m] + \sum_{l=1}^{l_{\max}} \bar{p}[L+1-l,m] \cdot \bar{f}[l-1,m] + \sum_{l_{\max}+1}^{L+1} \bar{p}[L+1-l,m] \cdot \bar{f}[l-1,m].$$

Using an index shift $l \to l + 1$ in the last sum and the fact that we can compute $\bar{f}[l,m]$ from $\bar{f}[l-1,m]$ if $l > l_{\max}$, we get

$$\bar{p}[L+1,m] = \bar{f}[L,m] + \sum_{l=1}^{l_{\max}} \bar{p}[L+1-l,m] \cdot \bar{f}[l-1,m] + \left(1 - \frac{1}{|\Sigma|}\right) \cdot \bar{p}[L - l_{\max},m] \cdot \bar{f}[l_{\max}-1,m]$$
$$+ \left(1 - \frac{1}{|\Sigma|}\right) \cdot \left(\sum_{l=l_{\max}+1}^{L} \bar{p}[L-l,m] \cdot \bar{f}[l-1,m]\right).$$

The first three terms take $\mathrm{O}(l_{\max})$ to compute, the last sum has already been computed for $\bar{p}[L,m]$ and is now available in $\mathrm{O}(1)$ if we stored it in that step or in $\mathrm{O}(l_{\max})$ if it has to be recomputed from $\bar{p}[L,m]$. □

We give two short examples on the size of these tables which also show that usually, the factor $l_{\max}$ is neglectable in applications.

Example 12. For applications such as protein identification usual parameters are $m_{\max} = 30\,000$ for $\Delta = 0.1$ and a maximal string length of $L_{\max} = 1000 \ldots 3000$. It is thus feasible to do exact computations and store p in memory. Using the amino acid alphabet, we have $\mu_{\min} \approx 500$, yielding a maximal detectable fragment length $l_{\max} \approx 60$.

Other applications such as bacteria identification from nucleic acids patterns [21] may require more than 10^8 entries using $m_{\max} \approx 100\,000$ for mass accuracy 0.1 Da, and $L_{\max} \approx 1000$. In this case, $\mu_{\min} \approx 2892$ and we get $l_{\max} \approx 35$.

To reduce memory consumption, we can leave out those rows $p[\cdot, m]$ where m has no decomposition as a fragment $y \in \Sigma_x^*$ with $\mu(y) = m$. Furthermore, we can usually discard columns $p[L, \cdot]$ where L is below a certain lower bound $L_{\min}$.

For the tryptic digestion of amino acids, the enzyme cleaves after the C-terminus of both lysine (K) and arginine (R) *except* before proline (P). We can capture this by a recurrence similar to (7) computable with the same time complexity, see [19] for details.

Example 13. We consider the alphabet $\Sigma := \{\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}\}$ with cleavage character $x := \mathrm{D}$, and masses $\mu(\mathrm{A}) = 3$, $\mu(\mathrm{B}) = 5$, and $\mu(\mathrm{C}) = 6$. Computation of $c[\cdot,\cdot]$ is straightforward, for example $c[5, 20] = c[4, 14] + c[4, 15] + c[4, 17] = 4 + 4 + 12 = 20$. For $m := 20$, the complete column $c[l, 20]$ reads:

l	0	1	2	3	4	5	6	7+
$c[l, 20]$	0	0	0	0	13	20	6	0

For computing $p[L, m]$ we use the recurrence of Theorem 10:

L	0	1	2	3	4	5	6	7	8	9	10
$p[L, m]$	0	0	0	0	$\frac{13}{256}$	$\frac{46}{1024}$	$\frac{163}{4096}$	$\frac{712}{16384}$	$\frac{3142}{65536}$	$\frac{13575}{262144}$	$\frac{58653}{1048576}$

So, $p[10, 20] = \frac{58653}{1048576} = 0.05593\ldots$ is the probability to draw a string $s \in \Sigma^{10}$ that generates a fragment of mass $m = 20$.

6. Significance of alignment scores

Using alignment scores as introduced above allows us to select a best-scoring simulated reference spectrum from, say, a database of sequences. But what are the chances that this score can be achieved by chance alone? Using the occurrence probabilities from Section 5, we can now compute the contribution of individual peaks p'_j of a sample spectrum $\mathcal{S}'$ to the total score. We will analyse the many-to-one matching scenario because it allows us to model alignment scores using only mild independence assumptions. We again consider the simple random model of random strings s with uniformly drawn characters. For better readability, we confine ourselves to peaks having mass as their only attribute and identify a peak p with its mass m, e.g. writing $\Psi_j(m)$ for $\Psi_j(p)$.

Recall that $p[L, m]$ is the probability that the random string s generates a fragment y with mass $\mu(y) = m$. Our goal is to define a random variable as the score of matching sample peak p'_j to all "adequate" peaks of the reference spectrum.

First, let X_j^{match} be the random variable that sums the scores over all peaks in the reference spectrum $\mathcal{S}$ that we can match with peak p'_j: For $m \in \mathcal{U}_j$ we use $S[L, m]$ from (5) and define

$$X_j^{\text{match}}(s) := \sum_{m \in \mathcal{U}_j} \Psi_j(m) \cdot \mathbb{1}\big(s \in S[L, m]\big)$$

where $\mathbb{1}(\cdot)$ denotes the indicator function. Assuming independence, we can easily see

$$\mathbb{E}\big(X_j^{\text{match}}\big) = \sum_{m \in \mathcal{U}_j} p[L, m] \cdot \Psi_j(m) \tag{8}$$

and

$$\operatorname{Var}\big(X_j^{\text{match}}\big) = \sum_{m \in \mathcal{U}_j} p[L,m] \cdot \Psi_j(m)^2 - \big(\mathbb{E}\big(X_j^{\text{match}}\big)\big)^2. \tag{9}$$

Second, if there is no peak in the reference spectrum with mass $m \in \mathcal{U}_j$, then the sample peak p'_j is an additional peak and must be penalized by adding $\Psi_j^{\text{add}} = \mathit{score}(\varepsilon, p'_j)$ to the spectrum score. We define the random variable

$$X_j^{\text{add}}(s) := \Psi_j^{\text{add}} \cdot \mathbb{1}\Bigg(s \notin \bigcup_{m \in \mathcal{U}_j} S[L,m]\Bigg)$$

for the "additional peak score". To simplify computations, we assume independence of the events that s generates fragments of distinct masses. Then, the probability that $X_j^{\text{add}}(s) = \Psi_j^{\text{add}}$ holds, is

$$\mathbb{P}\big(X_j^{\text{add}}(s) = \Psi_j^{\text{add}}\big) \approx \bar{p}_j := \prod_{m \in \mathcal{U}_j} \bar{p}[L,m]. \tag{10}$$

Now, the expected additional score of peak p'_j and its variance are given by

$$\mathbb{E}\big(X_j^{\text{add}}\big) = \bar{p}_j \cdot \Psi_j^{\text{add}}$$

and

$$\operatorname{Var}\big(X_j^{\text{add}}\big) = \bar{p}_j \cdot \big(\Psi_j^{\text{add}}\big)^2 - \big(\mathbb{E}\big(X_j^{\text{add}}\big)\big)^2.$$

We estimate expected value and variance of the random variable $X_j := X_j^{\text{match}} + X_j^{\text{add}}$ as

$$\mathbb{E}(X_j) = \mathbb{E}\big(X_j^{\text{match}}\big) + \mathbb{E}\big(X_j^{\text{add}}\big) \tag{11}$$

and

$$\operatorname{Var}(X_j) = \operatorname{Var}\big(X_j^{\text{match}}\big) + \operatorname{Var}\big(X_j^{\text{add}}\big) - 2\mathbb{E}\big(X_j^{\text{match}}\big) \cdot \mathbb{E}\big(X_j^{\text{add}}\big) \tag{12}$$

in view of

$$\operatorname{Cov}\big(X_j^{\text{match}}, X_j^{\text{add}}\big) = \mathbb{E}\big(X_j^{\text{match}} \cdot X_j^{\text{add}}\big) - \mathbb{E}\big(X_j^{\text{match}}\big) \cdot \mathbb{E}\big(X_j^{\text{add}}\big) = -\mathbb{E}\big(X_j^{\text{match}}\big) \cdot \mathbb{E}\big(X_j^{\text{add}}\big)$$

because either $X_j^{\text{match}}(s) = 0$ or $X_j^{\text{add}}(s) = 0$ holds for all $s \in S_{L,m}$.

Now, we consider peaks in the reference spectrum that we cannot match to a peak of the sample spectrum. Define $\mathcal{M}^+ := \bigcup_{j=1}^{n'} \mathcal{U}_j$ as the support of all peak scoring functions, then $\mathcal{M}^- := \mathcal{M} \setminus \mathcal{M}^+$ is the set of reference masses that cannot be matched with any sample peak. Any reference peak p with mass $m \in \mathcal{M}^-$ is therefore a missing peak, and must be penalized by $\Psi^{\text{miss}}(m)$. We define random variables X_m^{miss} for $m \in \mathcal{M}^-$ by $X_m^{\text{miss}}(s) := \Psi^{\text{miss}}(m)$ if the reference string s generates a fragment of mass m, and $X_m^{\text{miss}}(s) := 0$ otherwise. We easily calculate

$$\mathbb{E}\big(X_m^{\text{miss}}\big) = p[L,m] \cdot \Psi^{\text{miss}}(m) \tag{13}$$

and

$$\operatorname{Var}\big(X_m^{\text{miss}}\big) = p[L,m] \cdot \Psi^{\text{miss}}(m)^2 - \big(\mathbb{E}\big(X_m^{\text{miss}}\big)\big)^2. \tag{14}$$

We can compute $\sum_{m \in \mathcal{M}} p[L,m] \cdot \Psi^{\text{miss}}(m)$ and $\sum_{m \in \mathcal{M}} p[L,m] \cdot \Psi^{\text{miss}}(m)^2$ during preprocessing, what allows us to limit computations to masses $m \in \mathcal{M}^+$ in (16).

Finally, the random variable X is the total score of aligning the reference spectrum of a string $s \in \Sigma^L$ to the sample mass spectrum $\mathcal{S}' = \{p'_1, \ldots, p'_{n'}\}$, see (4). From the above,

$$X = \sum_{j=1}^{n'} X_j + \sum_{m \in \mathcal{M}^-} X_m^{\text{miss}} \tag{15}$$

and if we assume that these random variables are independent, we infer the expected score of sample spectrum $\mathcal{S}'$:

$$\mathbb{E}(X) = \sum_{j=1}^{n'} \mathbb{E}(X_j) + \sum_{m \in \mathcal{M}^-} \mathbb{E}\big(X_m^{\text{miss}}\big) \tag{16}$$

and its variance as

$$\operatorname{Var}(X) = \sum_{j=1}^{n'} \operatorname{Var}(X_j) + \sum_{m \in \mathcal{M}^-} \operatorname{Var}\big(X_m^{\text{miss}}\big). \tag{17}$$

Also, X is the sum of many nearly independent random variables, so X can be approximated by a *normal distribution* $\mathcal{N}(\bar{\mu}, \bar{\sigma}^2)$ with mean $\bar{\mu} := \mathbb{E}(X)$ and variance $\bar{\sigma}^2 := \operatorname{Var}(X)$ using the central limit theorem.

In the above calculations, we had to assume independence of random variables though these variables are slightly correlated. To show that our estimations are accurate in application settings we have performed simulations, see Section 8.

Suppose we have computed an alignment score $sc := score(\mathcal{S}, \mathcal{S}')$ for a sample mass spectrum $\mathcal{S}'$ and a reference mass spectrum $\mathcal{S}$ generated *in silico* from a string s. This was done using either the simple many-to-one alignment of Section 4, or the more elaborate merging alignment. This score is now a realization of the random variable X as defined above using $\mathcal{S}'$ and the length L of the sample string s from which the reference spectrum $\mathcal{S}$ was derived. We can compute the expectation $\bar{\mu}$ and variance $\bar{\sigma}^2$ of X in constant space and $\mathrm{O}(|\mathcal{M}_+|) = \mathrm{O}(|\mathcal{S}'| \cdot u)$ time, where $u = \max_j |\mathcal{U}_j|$ is the maximal width of any support. We can then compute the p-value of sc using $Z \sim \mathcal{N}(0, 1)$ and the equation

$$\mathbb{P}(X \geqslant sc) \approx \mathbb{P}\left(Z \geqslant \frac{sc - \bar{\mu}}{\bar{\sigma}}\right). \tag{18}$$

7. Collision-induced dissociation of random strings

So far, we did only consider statistics of fragments resulting from character-specific cleavage. Let us now focus on the second important technique to identify proteins by mass spectrometry: Collision induced dissociation of peptides by tandem mass spectrometry. Here, we break the peptide string s of known *parent mass* M into all prefixes and suffixes of s. We concentrate on the main ion series (b/y-ions) and ignore mass modifications of b/y-ions (addition H group for b-ions, additional H_3O group for y-ions) for the sake of readability. Again, we can easily incorporate these mass modifications as well as other ion series.

We require all masses to be natural numbers. The peaks detected in the sample mass spectrum correspond to prefixes and suffixes of the amino acid string s. Regarding s, we know its *parent mass* $M := \mu(s)$. So, we want to uniformly sample from the set $S[M] := \{s \in \Sigma^*\colon\ \mu(s) = M\}$. What is the probability that any such string has a prefix or suffix y of mass $\mu(y) = m$? Recall that we can easily compute the number of strings $s \in S[M]$ with $\mu(s) = m$ using (6).

The surprisingly simple result of this section is:

Theorem 14. *Let $d[m]$ denote the number of strings $s \in \Sigma^*$ with $\mu(s) = m$. For parent mass $M \in \mathbb{N}$, let s be a string uniformly drawn from the set of strings $S[M]$ with mass M. The probability that s has a prefix of mass $m \in \{1, \ldots, M\}$, is*

$$\tilde{q}[M, m] := \frac{1}{d[M]} d[m] d[M - m]. \tag{19}$$

Set $\bar{m} := \min\{m, M - m\}$, then the probability that s has a prefix or suffix of mass m, is

$$q[M, m] := \frac{1}{d[M]} \big(2d[\bar{m}]d[M - \bar{m}] - d[\bar{m}]^2 d[M - 2\bar{m}]\big). \tag{20}$$

Theorem 14 allows us to compute $q[M, m]$ in constant time, if $d[\cdot]$ is known.

For fixed prefix mass m, let $\tilde{b}[M]$ denote the number of strings in $S[M]$ of parent mass $M \in \mathbb{N}$ that have a prefix y of mass $\mu(y) = m$. Analogously, let $b[M]$ denote the number of strings in $S[M]$ that have a prefix of mass m_1, or a prefix of mass m_2, for fixed prefix masses $m_1 \leqslant m_2$. The proof of Theorem 14 is based on the following observations:

Lemma 15. *Let $M \in \mathbb{N}$ denote the parent mass. If $m \leqslant M$ is the forced prefix mass, then*

$$\tilde{b}[M] = d[m]d[M-m].$$

If $m_1 \leqslant m_2 \leqslant M$ are the forced prefix masses, then

$$b[M] = d[m_1]d[M-m_1] + d[m_2]d[M-m_2] - d[m_1]d[m_2-m_1]d[M-m_2].$$

Proof. First, we ask for the number of strings $s \in \Sigma^*$ with prefix mass m and parent mass M. This implies that s is of the form $s = yz$ for strings $y, z \in \Sigma^*$ with $\mu(y) = m$ and $\mu(z) = M - m$. As we can combine any two such prefix/suffix strings, and since there exist $d[m]$ such prefixes and $d[M-m]$ such suffixes, we conclude $\tilde{b}[M] = d[m]d[M-m]$.

The second part of the lemma follows by inclusion/exclusion arguments: From the previous, we know that there exist $d[m_1]d[M-m_1]$ strings with prefix mass m_1, and $d[m_2]d[M-m_2]$ strings with prefix mass m_2. From this, we have to subtract the number of strings that contain both m_1 and m_2 as a prefix mass, and this number is $d[m_1]d[m_2 - m_1]d[M-m_2]$. □

The proof of Theorem 14 then follows immediately from the definitions.

Example 16. Set $\Sigma := \{\mathrm{A}, \mathrm{B}, \mathrm{C}\}$, $\mu(\mathrm{A}) = 3$, $\mu(\mathrm{B}) = 4$, and $\mu(\mathrm{C}) = 6$. Let $m := 8$ be the prefix/suffix mass. We compute the table d as follows: We can compute $\tilde{b}$ and b for $m_1 := m$ and $m_2 := M - m$ by a recurrence similar to that for computing $d[\cdot]$: We use recurrence relation (6) on $\tilde{b}, b$ except that we force $\tilde{b}[0] := b[0] := 0$, $\tilde{b}[m] := b[m] := d[m]$, and $b[M-m] := d[M-m]$. For string mass $M \leqslant 20$ we get (columns containing only zeros not shown):

M	0	3	4	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
$d[\cdot]$	1	1	1	2	2	1	3	5	3	6	10	9	12	21	22	27	43	52
$\tilde{b}[\cdot]$	0	0	0	0	0	1	0	0	1	1	0	2	2	1	3	5	3	6
$b[\cdot]$	0	0	0	0	0	**1**	0	0	1	**6**	0	2	7	6	3	15	13	11

Using Lemma 15 we calculate $\tilde{b}[20] = d[8]d[20-8] = 1 \cdot 6 = 6$ as expected. The probability to draw a string $s \in S[20]$ with prefix or suffix of mass $m = 8$, is

$$q[20, 8] = \frac{1}{d[20]}\bigl(2d[8]d[12] - d[8]^2 d[4]\bigr) = \frac{1}{52}(2 \cdot 1 \cdot 6 - 1^2 \cdot 1) = \frac{11}{52} = 0.2115\ldots.$$

We want to stress that we *cannot* estimate $q[M, m]$ by $2\tilde{q}[M, m] - \tilde{q}[M, m]^2$, assuming independence of events: For example, let $M := 20$ and $m := 10$, then this estimation gets $2\frac{5 \cdot 5}{52} - \frac{25 \cdot 25}{52 \cdot 52} = 0.7303\ldots$. But every string s with $\mu(s) = 20$ has a prefix y with $\mu(y) = 10$ if and only if it has a suffix y' with $\mu(y') = 10$! So, these events are completely dependent, and we find $q[20, 10] = \tilde{q}[20, 10] = \frac{25}{52} = 0.4807\ldots$.

Analogously to Section 5, we define random variables X_j^{match}, X_j^{add}, and X_m^{miss} to estimate mean and variance of alignment scores. Here the set of relevant strings is

$$S[M, m] := \bigl\{s \in S[M]\colon s \text{ has prefix or suffix of mass } m\bigr\}.$$

Usually, estimation of mean and variance is analogous to Section 5 replacing $p[L, m]$ by $q[M, m]$, see (8) and (13). The major difference is computation of $\bar{p}_j$ for the random variable X_j^{add} in (10): We may safely expect that $|\mathcal{U}_j| < \mu_{\min}$ is smaller than the minimal mass of any character. This implies that some string cannot have two prefixes (or two suffixes) with masses both in the support $\mathcal{U}_j$. Assuming disjointness of the remaining cases we reach

$$\mathbb{P}\bigl(X_j^{\text{add}}(s) = \Psi_j^{\text{add}}\bigr) \approx \bar{p}_j := 1 - \sum_{m = \mathcal{U}_j} q[M, m].$$

Exact calculation of $\bar{p}$ requires $O(|\mathcal{U}_j|^2)$ time, using a generalization of Lemma 15.

Finally, the random variable $X = \sum_j X_j + \sum_{m \in \mathcal{M}^-} X_m^{\text{miss}}$ is the total score of aligning the reference spectrum of a string $s \in S[M]$ to the sample mass spectrum $\mathcal{S}'$. Again, we must assume that these random variables are independent, what is less correct than in the previous section, because peaks with mass differences of character masses are clearly correlated. Simulations on how to correct the resulting skew of the estimated variance are currently executed. The remaining calculations are analogous to Section 5, and we can compute the p-value of the alignment score.

8. Results

We want to asses the quality of our estimations for tryptic digestion of amino acids as described above. We use integer masses with accuracy $\Delta = 0.1$, and the following scoring scheme: Additional and missing peaks are penalized with score -0.2, matched peaks are given the Gaussian score described in Example 8 using a standard deviation of 2 Da and a threshold of 95%. We do the following for $L = 350, 500$: We draw a random sample string of length L and simulate its cleavage pattern under tryptic digestion. Then, we draw 250 000 random strings of length L and compute the alignment score for the respective mass spectra. Finally, we estimate mean and variance of a normal distribution using the method of Section 6. To demonstrate the correctness of the normal distribution assumption, normal-quantile–

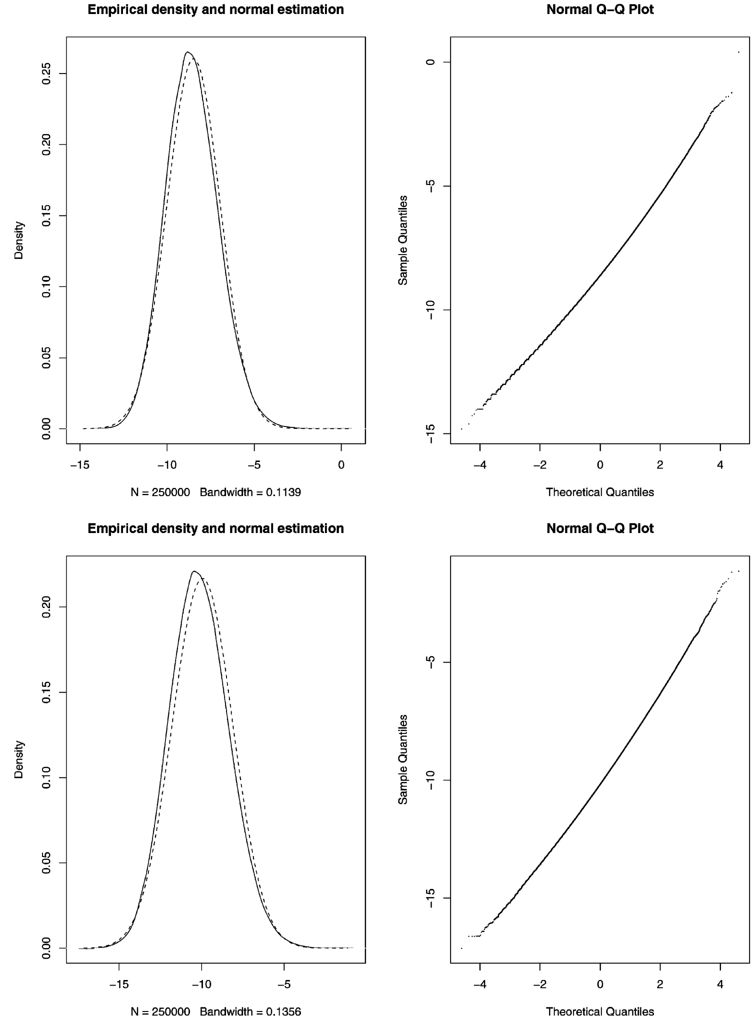

Fig. 1. Simulation results: Distribution of scores for randomly drawn strings (solid line) and normal distribution estimated using the method of Section 5 (dashed line) as well as quantile–quantile-plot for normal distribution for length $L = 350$ (top) and $L = 500$ (bottom).

quantile-plots are provided for each string-length, showing the quantiles of the predicted distribution against the quantiles of a normal distribution (see e.g. [22] for details); a straight line would indicate a perfect agreement between the two distributions. The plots in Fig. 1 clearly show that our approach allows quite accurate estimation of the distribution of scores.

9. Discussion and improvements

We presented a general approach for aligning two mass spectra and, in particular, aligning a sample spectrum and a (theoretically predicted) reference spectrum. Our approach allows very general and flexible scoring schemes that may take into account not only masses, but also arbitrary other peak attributes such as intensity or area-under-curve. This approach also allows un-symmetric scores, so we can score measured intensities, which currently cannot be predicted from sequence.

To assess the significance of an alignment score, we showed that the score distribution, unlike sequence alignment scores, can be approximated by a normal distribution. We gave a general approach to compute the moments, in particular the expectation and variance, of the score distribution in linear time, which allows to estimate the p-value of the score.

Regarding substrings of certain mass in a random string, we presented efficient methods to compute occurrence probabilities. We believe that this approach will allow for generalizations to related questions in the context of weighted strings.

We are currently evaluating spectrum alignments of tryptic digestion data using protein databases. In particular, scoring schemes are tested for their discriminative power and compared to existing approaches such as MASCOT. We will also integrate our algorithms into the in-house ProDB system [23].

Acknowledgements

Sebastian Böcker was supported by "Deutsche Forschungsgemeinschaft" (BO 1910/1-1) within the Computer Science Action Program. Hans-Michael Kaltenbach is supported by the "International NRW Graduate School in Bioinformatics and Genome Research". The authors want to thank Sven Rahmann for helpful discussions.

References

[1] R. Aebersold, M. Mann, Mass spectrometry-based proteomics, Nature 422 (2003) 198–207.

[2] S.D. Patterson, R. Aebersold, Mass spectrometric approaches for the identification of gel-separated proteins, Electrophoresis 16 (1995) 1791–1814.

[3] R.G. Cooks (Ed.), Collision Spectroscopy, Plenum Press, New York, 1978.

[4] B. Lu, T. Chen, A suboptimal algorithm for de novo peptide sequencing via tandem mass spectrometry, J. Comput. Biol. 10 (1) (2003) 1–12.

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

[6] I.-J. Wang, C.P. Diehl, F.J. Pineda, A statistical model of proteolytic digestion, in: Proceedings of IEEE CSB 2003, Stanford, CA, 2003, pp. 506–508.

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

[8] X. Huang, M.S. Waterman, Dynamic programming algorithms for restriction map comparison, Comput. Appl. Biosci. 8 (5) (1992) 511–520.

[9] H. Hermjakob, R. Giegerich, W. Arnold, RIFLE: Rapid identification of microorganisms by fragment length evaluation, in: Proceedings of ISMB 1997, Halkidiki, Greece, 1997, pp. 131–139.

[10] T. Aittokallio, P. Ojala, T.J. Nevalainen, O. Nevalainen, Automated detection of differently expressed fragments in mRNA differential display, Electrophoresis 22 (10) (2001) 1935–1945.

[11] C. Wenk, Applying an edit distance to the matching of tree ring sequences in dendrochronology, in: M. Crochemore, M. Paterson (Eds.), Proceedings of Combinatorial Pattern Matching (CPM99), in: Lecture Notes in Computer Science, vol. 1645, 1999, pp. 223–242.

[12] V. Mäkinen, Peak alignment using restricted edit distances, Biomolecular Engineering, submitted for publication.

[13] P.A. Pevzner, V. Danćik, C.L. Tang, Mutation-tolerant protein identification by mass spectrometry, J. Comput. Biol. 7 (6) (2000) 777–787.

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

[15] J. Colinge, A. Masselot, J. Magnin, A systematic statistical analysis of ion trap tandem mass spectra in view of peptide scoring, in: Proc. of WABI 2003, Budapest, Hungary, in: Lecture Notes in Computer Science, vol. 2812, Springer, 2003, pp. 25–38.

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

[17] M. Karas, F. Hillenkamp, Laser desorption ionization of proteins with molecular masses exceeding 10,000 Daltons, Anal. Chem. 60 (1988) 2299–2301.

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

[19] H.-M. Kaltenbach, H. Sudek, S. Böcker, S. Rahmann, Statistics of cleavage fragments in random weighted strings, Tech. Rep. 2005-06, Technische Fakultät der Universität Bielefeld, Abteilung Informationstechnik, 2005, http://bieson.ub.uni-bielefeld.de/volltexte/2006/900/.

[20] V. Danćik, T.A. Addona, K.R. Clauser, J.E. Vath, P.A. Pevzner, De novo peptide sequencing via tandem mass spectrometry, J. Comput. Biol. 6 (3/4) (1999) 327–342.

[21] F. von Wintzingerode, S. Böcker, C. Schlötelburg, N.H. Chiu, N. Storm, C. Jurinke, C.R. Cantor, U.B. Göbel, D. van den Boom, Base-specific fragmentation of amplified 16S rRNA genes and mass spectrometry analysis: A novel tool for rapid bacterial identification, Proc. Natl. Acad. Sci. USA 99 (10) (2002) 7039–7044.

[22] R.A. Becker, J.M. Chambers, A.R. Wilks, The New S Language, Wadsworth & Brooks, 1988.

[23] A. Wilke, C. Rückert, D. Bartels, M. Dondrup, A. Goesmann, A.T. Hüser, S. Kespohl, B. Linke, M. Mahne, A.C. McHardy, A. Pühler, F. Meyer, Bioinformatics support for high-throughput proteomics, J. Biotechnol. 106 (2–3) (2003) 147–156.