BMC Bioinformatics

BioMed Central

Research article

Open Access

SAMPI: Protein Identification with Mass Spectra Alignments

Hans-Michael Kaltenbach*[1], Andreas Wilke[2] and Sebastian Böcker*[3]

Address: [1]AG Genominformatik, Technische Fakultät, Universität Bielefeld, PF 100 131, 33501 Bielefeld, Germany, [2]Computation Institute, University of Chicago, Chicago, IL 60637, USA and [3]Lehrstuhl für Bioinformatik, Friedrich-Schiller-Universität Jena, Ernst-Abbe-Platz 2, 07743 Jena, Germany

Email: Hans-Michael Kaltenbach* - michael@cebitec.uni-bielefeld.de; Andreas Wilke - wilke@mcs.anl.gov; Sebastian Böcker* - boecker@minet.uni-jena.de

* Corresponding authors

Published: 26 March 2007

BMC Bioinformatics 2007, **8**:102 doi:10.1186/1471-2105-8-102

Received: 5 July 2006
Accepted: 26 March 2007

This article is available from: http://www.biomedcentral.com/1471-2105/8/102

Abstract

Background: Mass spectrometry based peptide mass fingerprints (PMFs) offer a fast, efficient, and robust method for protein identification. A protein is digested (usually by trypsin) and its mass spectrum is compared to simulated spectra for protein sequences in a database. However, existing tools for analyzing PMFs often suffer from missing or heuristic analysis of the significance of search results and insufficient handling of missing and additional peaks.

Results: We present an unified framework for analyzing Peptide Mass Fingerprints that offers a number of advantages over existing methods: First, comparison of mass spectra is based on a scoring function that can be custom-designed for certain applications and explicitly takes missing and additional peaks into account. The method is able to simulate almost every additive scoring scheme. Second, we present an efficient deterministic method for assessing the significance of a protein hit, independent of the underlying scoring function and sequence database. We prove the applicability of our approach using biological mass spectrometry data and compare our results to the standard software Mascot.

Conclusion: The proposed framework for analyzing Peptide Mass Fingerprints shows performance comparable to Mascot on small peak lists. Introducing more noise peaks, we are able to keep identification rates at a similar level by using the flexibility introduced by scoring schemes.

Background

Protein identification using mass spectrometry has become one of the central tools in proteomics and systems biology [1]: With growing protein sequence databases such as SwissProt [2], fast and accurate identification of a sample protein remains a central problem. There are two common strategies for protein identification using mass spectrometry: Peptide Mass Fingerprints [3] and protein identification from peptide sequence information using tandem mass spectrometry [4].

Peptide mass fingerprinting (PMF) is preceded by a protein separation step using gel or chromatographic separation. The separated protein is digested by specific enzymatic cleavage such as tryptic digestion, followed by mass spectrometric measurement of the resulting peptides. The resulting mass spectrum has to be preprocessed into a list of signal peaks that form the input to identification algorithms. In our approach, we concentrate on Matrix Assisted Laser Desorption/Ionization (MALDI) [5], the predominant ionization technique for PMF. This technique produces mainly singly charged ions, allowing us to talk of the mass *m* of a molecule, instead of its mass-to-charge ratio *m*/*z*.

To identify a measured protein from a sequence database, the database sequences are digested in-silico and each predicted peak list is matched and scored with the measured peak list. Usually, computation of the statistical significance should follow, using a statistical background model. Software routinely used for identification of proteins using PMF includes the commercial systems Mascot [6] which uses peak counting together with heuristic information and ProFound [7] which relies on a bayesian scoring scheme. These systems have a comparable performance [8].

In [9], we presented a new approach for PMF protein identification. The approach is based on a re-formulation of the identification problem as a global alignment problem. Further, p-values of identifications are computed using a combinatorial algorithm using uniform character frequencies.

The statistics for p-value computation is extended to a broader class of digestion enzymes and to arbitrary protein sequence models of independently and identically distributed (i.i.d.) amino acids in [10], where this model is also shown to be consistent with corresponding empirical SwissProt data.

Here, we validate the theoretical approach of peak list alignments as introduced in [9] and show the applicability of this approach on real proteomics data. We discuss several aspects of general scoring schemes to be used in peak list alignments; such schemes provide a unified framework that allows emulation and combination of already existing methods and ideas. We demonstrate how missing and additional peaks can explicitly be taken into account and peak intensities can be consistently added into the scoring procedure. We evaluate our method, called *SAMPI* (SAMPI: aligning mass spectra for protein identification), on real proteomics mass spectrometry data and compare the method to PMF identification using the standard software Mascot.

Results and discussion

To evaluate our method, 375 PMF tryptic mass fingerprints of charge state $[M + H]^+$ from an in-house proteomics experiment on the organism *Corynebacterium glutamicum* (Cg) are measured on a Bruker Ultraflex mass spectrometer. The proteins are separated using SDS-PAGE gel electrophoresis before mass measurement. Well-separated spots are digested with trypsin and peptide masses measured by mass spectrometry. For identification, trypsin is set as a cleavage enzyme, Carbamidomethyl is set as a fixed mass modification of ≈ +57 Da for Cysteine, a mass tolerance of 1 Da is set, and no missed cleavages are allowed.

Processing the raw spectra

To assess robustness and flexibility of the method, two different peak lists are derived for each raw spectrum. First, the peak list from the manufacturers peak detection software: It is conservative in picking only the highest abundant peaks; with about 0–90 peaks, 20 on average, these peak lists were comparatively small. Second, we apply a peak detection algorithm developed in our group that derives much larger peak lists of 34–729 peaks, 277 on average. A comparison of the peak list lengths is shown in Figure 1. For unknown reasons, the manufacturers software only delivers 325 peak lists, 9 of which were empty. The other algorithm delivers 375 valid peak lists. For better comparison, we differentiate the peak lists delivered by the algorithm of our group in the following by "PL" and "PL_{316}", denoting the whole set of peak lists and the set of peak lists where the manufacturers peak detection also delivers a corresponding nonempty peak list. Due to the different peak detection, the mass ranges for the measured and predicted peak lists were set to 500–3000 Da for the Bruker software, and 800–3000 Da for our peak lists. All peaks outside this range are discarded.

Databases

Both sets of peak lists are identified using Mascot versions v1.9 and v2.1 and the Gaussian scoring scheme described below with different parameters.

For estimating the false positive rate, we proceed as follows: In a first step, all peak lists are identified using the in-house Cg protein sequence database with 3,510

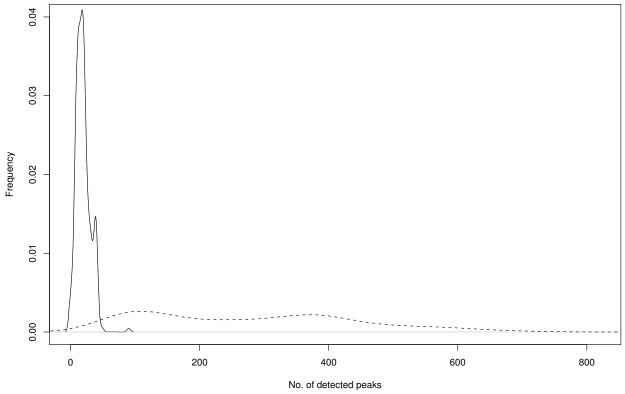

Figure 1
Peaklist sizes: Number of detected peaks using the Bruker software (solid) and our peak detection method (dashed).

sequences. The sequence identifiers for each identification are recorded for later comparison.

In a second step, all peak lists are again identified using a concatenation of the SwissProt database, release 48 with 155,824 entries, to the above Cg database. An identification is assumed to be correct if it is the same Cg sequence as recorded before. Conversely, an identification is assumed to be incorrect, if it is not a Cg sequence.

This approach makes it necessary to discard protein sequences that are equal to or highly similar to any Cg sequence from the SwissProt database beforehand. We therefore do an all-against-all comparison of the SwissProt database with 194,317 sequences and the Cg database with BLAST [11]. Sequences from the SwissProt database with an e-value of 10^{-30} or better are discarded and the remaining 155,824 sequences are appended to the Cg database.

The performance of PMF identification algorithms can thus be evaluated and compared independently on the set of sample mass spectra.

Results

The runtimes of both SAMPI and Mascot on the modified SwissProt database (FASTA format, no database indexing) and each of the two peak lists including all preprocessing times are listed in Table 1. The results for both versions of Mascot and the Gaussian scheme with several additional/missing scores are listed in Table 2. All parameter combinations were tested with and without use of peak intensities. Not surprising, different parameter sets lead to different numbers of correct identifications. Nevertheless, these numbers do not change rapidly with changing parameters, indicating a robust behavior of the alignment identification procedure. Using the small manufacturer's peak lists, a small penalty of additional and missing peaks yields a comparable number of correct identifications as Mascot. Using peak intensities in the scoring, this number drops considerably. This is most likely due to the fact that these peak lists already consist of the highest abundant peaks, which are now scaled to 1/3 to 1, distorting the relevance of peaks. This problem might be resolved by using a full rank statistic to scale peak intensities as used, e.g., in [12], instead of the implemented robust linear rescaling. Since we describe a proof-of-concept of the peak list alignment framework, we did not investigate this issue further. Using the larger, noisy peak lists results in the complete opposite behavior: Now, without using intensities to discriminate important and non-important peaks, the identification rate drops to about 1/2 for both the Gaussian schemes and Mascot. Additionally using the robust intensities leads to a good identification rate again. Note that now, higher penalties for additional and missing peaks are also helpful.

Table 1: Runtimes

	Bruker	PL
SAMPI	23 min (4.2 sec)	99 min (15.8 sec)
Mascot	43.5 min (8 sec)	192 min (30.7 sec)

Runtimes for identification of 325 small (Bruker, Bruk.) and 375 large (PL) peak lists with an average of 20 and 277 peaks, respectively. Values in parentheses are runtimes per spectrum.

Table 2: Identification results

		w/out intensity			w/intensity		
		Bruk.	PL	PL_{316}	Bruk.	PL	PL_{316}
Mascot v1.9		123	58	53	-	-	-
Mascot v2.1		119	59	53	-	-	-
SAMPI							
c_1	c_2						
-0.1	-0.1	112	56	51	72	106	87
-0.2	-0.2	111	56	51	78	96	92
-0.3	-0.3	96	54	48	65	103	98
-0.4	-0.3	89	53	48	52	110	105
-0.4	-0.4	91	54	49	54	108	103
-0.4	-0.5	94	53	48	57	109	104

Number of correctly identified spectra in the Cg+SwissProt database, using Gaussian score, missing/additional peak penalties shown in first two columns. Three different peak lists are considered: Bruk: The peak lists by the Bruker software, PL and PL_{316} which refer to the peak lists by our detection algorithm where PL_{316} are restricted to the 316 out of 325 peak lists that the Bruker software also detected.

We found the score separation of correct and incorrect identifications to be comparable to Mascot (Figures 2 and 3).

Conclusion

We propose a new formulation of protein identification using Peptide Mass Fingerprinting as an *alignment problem*. We introduce general peak-wise *scoring schemes* and show how these can be used to score two peak lists by dynamic programming. The scoring schemes provide a large amount of flexibility by allowing the user to independently set matching, additional and missing scores. They also allow consistent inclusion of additional features such as peak intensities into the identification process.

A mathematical model based on *random weighted strings* is used to efficiently estimate the statistical significance of an alignment score. This model only needs character frequencies as parameters, which can be estimated even with small sequence databases, allowing the use of species-specific protein data. The *significance score* is then computed to get comparable scores independent of the underlying database and sequence lengths. We also propose a first

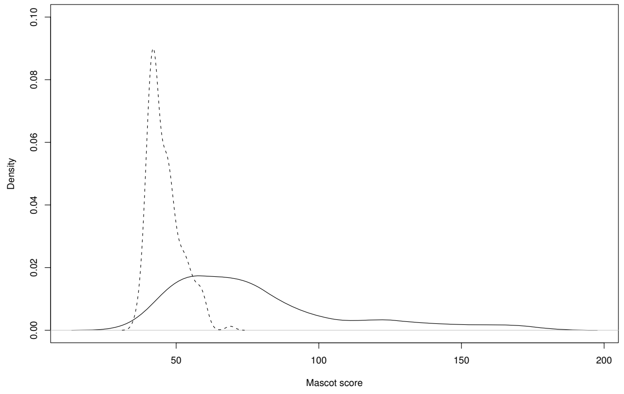

Figure 2
Score distributions. *Mascot*: Distribution of Mascot scores of correct (solid line) and incorrect (dashed line) identifications using the Cg+SwissProt database, 1 Da mass tolerance, no missed cleavages, and 316 spectra.

example of an alignment scoring scheme, called *Gaussian score*, using mass difference and robust intensities. We tested our approach on biological PMF data and compared our results to the standard software Mascot. We were able to correctly identify a comparable number of proteins using peak lists produced by the machine vendor's peak detection software. Using our own peak detection with about 8–10 times as many peaks, we showed the flexibility of our approach by correctly identifying approximately the same number of proteins whereas the performance of Mascot dropped considerably to about half the number of correct identifications.

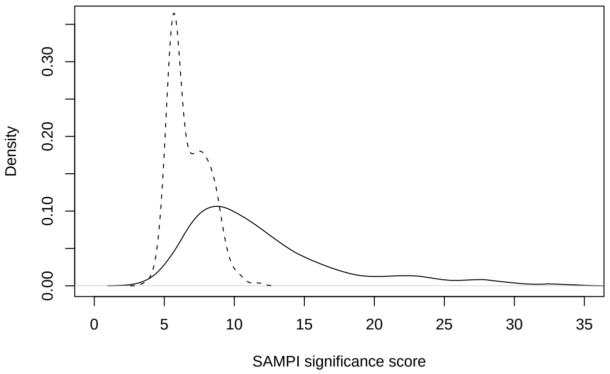

Figure 3
Score distributions. *SAMPI*: Distribution of SAMPI scores of correct (solid line) and incorrect (dashed line) identifications using the Cg+SwissProt database, parameter set B, no missed cleavages, and 316 spectra.

For the near future, we plan to incorporate missed cleavages into the program. They can easily be handled by the alignment algorithm, but the statistical model has to be extended slightly. Further, we plan to incorporate the method into the ProDB proteomics platform [13] and to compare it to other protein identification tools besides Mascot. As already discussed in the Results and Discussion section, the incorporation of intensity information is not optimal. This is not due to the framework but rather due to the use of scaled intensity values. Nevertheless, using full rank statistics or other probabilistic intensity incorporations are to be investigated in the future. As a last point, the score normalization is not as good as expected, especially on smaller peak lists; the reasons and possible improvements are to be investigated. The flexible alignment framework together with the deterministic, model-based significance computation seems promising, although some improvements are clearly necessary.

Methods

To identify a measured peak list using peak lists predicted from database sequences, a measure is needed for the similarity of two peak lists. Our scoring of similarity is based on a peak-wise scoring function to score a pair of peaks, one of them possibly being a "gap" peak. The optimal matching of two peak lists can then be computed in a way similar to global sequence alignment.

For computing a statistical significance of an alignment score, we introduce a null-model based on a random protein model and estimate the alignment score distribution.

The general method works as follows, where the individual steps are explained in detail below: In a pre-processing step, a peak list is computed from each entry in the particular protein sequence database; it is called the *predicted peak list* of the sequence. Further, several statistics are computed for later use in the identification's significance estimation. These statistics are the length distribution and the joint length-mass distribution of cleavage fragments. From these, the occurrence probabilities are computed for each possible fragment mass and each protein length contained in the database. Now, the highest scoring protein sequence is computed for each measured spectrum by aligning this spectrum to each predicted spectrum, computing the alignment score and returning the sequence with the highest scoring predicted spectrum. Further, a statistical significance is computed for each such alignment score.

Spectra alignments

Peaks and peak lists

Every peak p_i has a mass $m_i \in \mathcal{M}$ and possibly other attributes $(a_{i,1}, ..., a_{i,k}) \in \mathcal{A}$, $k \geq 0$. A *peak list* $\mathcal{S}$ of length n is a list $\mathcal{S} = \{p_1, ..., p_n\}$ of *peaks* $p_i \in \mathcal{M} \times \mathcal{A}$. A peak list is sorted by mass, thus $m_i < m_j$ if $i < j$.

Note that we allow the set $\mathcal{A}$ of peak attributes to be empty. The simplest type of peak is a peak having only its mass $m \in \mathbb{R}$. A peak with mass and relative intensity could be represented as an element of $\mathbb{R} \times [0,1]$.

Scoring peaks and spectra

Let $\mathcal{S}_p = \{p_1, ..., p_n\}$ and $\mathcal{S}_m = \{p'_1, ..., p'_{n'}\}$ be two peak lists of length n and n', respectively. For $1 \leq i \leq n$ and $1 \leq j \leq n'$, let $p_i \in \mathcal{M} \times \mathcal{A}$ and $p'_j \in \mathcal{M} \times \mathcal{A}'$ where we allow the sets of additional peak attributes $\mathcal{A}$ and $\mathcal{A}'$ to be different. An example would be a measured peak list $\mathcal{S}_m$, with relative intensity and a predicted peak list $\mathcal{S}_p$ with the generating string fragment as additional attributes. To compute an optimal matching between peak list $\mathcal{S}_p$ and peak list $\mathcal{S}_m$, we first define a scoring function that scores two individual peaks.

A *peak-wise scoring function score* is a function

$$\mathit{score}: (\mathcal{S}_p \cup \{\varepsilon\}) \times (\mathcal{S}_m \cup \{\varepsilon\}) \rightarrow \mathbb{R}$$

mapping a predicted and a measured peak to a real value. Here, ε denotes a special "gap" peak. For two peaks $p \in \mathcal{S}_p$ and $p' \in \mathcal{S}_m$, we say that *score*(p, p') is a *matching score*. We call *score*(p, ε) a *missing score* and p a *missing peak*, as it is not matched to any peak in $\mathcal{S}_m$. Similarly, *score*(ε, p') is called an *additional score* for an *additional peak* p'. For completeness, we define *score*$(\varepsilon, \varepsilon) := -\infty$.

We want to stress that missing and additional peaks are likely to be seen even if the measured spectrum stems from a measurement of a known sequence. Additional peaks, peaks that are seen in the measurement but cannot be explained by the sequence, may simply be chemical noise from the biochemical sample preparation. Missing peaks may occur due to incorrect peak detection or failed ionization of the corresponding fragment. Of course, missing and additional peaks also occur if the measured spectrum does not stem from the sequence under investigation.

Example 1 (Peak counting)

Using only peak mass as attribute, a peak counting score *could ignore missing and additional peaks, i.e. set score*(p, ε) = *score*(ε, p') = 0, *and give a positive score whenever the difference of the two peak masses m and m' is not too large:*

$$score(p, p') = \begin{cases} 1, & \text{if } |m - m'| \leq \delta, \\ -\infty, & \text{else} \end{cases}$$

for some positive constant δ.

Noting that it would be meaningless to match two pairs of peaks that overcross in mass, we compute the optimal matching between two spectra, i.e. the matching yielding the highest sum of peak-wise scores, as a global alignment, using the well-known dynamic programming recurrence. Let $E[i, j]$ denote the score for the optimal matching between the two spectra up to peaks p_i and p'_j, respectively. Then the alignment table is computed as

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

The score *score*$(\mathcal{S}_p, \mathcal{S}_m)$ of the optimal matching, given in $E[n, n']$, is called the *alignment score* of the spectra $\mathcal{S}_p$ and $\mathcal{S}_m$.

As in the case of sequence alignment, the optimal matching itself can be recovered by backtracking in the dynamic programming table $E[\cdot, \cdot]$. The alignment score can be computed in time $O(n \cdot n')$, but faster implementations are possible, using only a diagonal band in $E[\cdot, \cdot]$.

This approach is a standard technique [14,15], and has been successfully applied to such diverse problems as tree ring and liquid chromatography matching [16,17]. A more formal model of peak list alignment can be found in [9].

Scoring schemes

Although a peak in a measured peak list is described at least by its mass and absolute intensity, most identification algorithms only make use of its mass [18]. This is partly because mass is the most discriminative parameter measured and partly because intensity depends heavily on the actual parameter settings of the machine. The basis for many schemes is the observation that a measurement error between the "real" mass of a molecule and the measured mass can be described by a Gaussian distribution with mean 0 and a standard deviation *sd* dependent on the machine settings and experiment type. The mean might also deviate from 0 if the machine is not calibrated correctly. We will now introduce a family of scoring schemes, the *Gaussian scores*, that will be used in further sections to demonstrate the applicability of our approach. Note however, that the approach is by no means limited to this mass measurement error distribution.

Mass difference

The matching score *score*(p, p') for two peaks p and p' with masses m and m', respectively, is the probability of a Gaussian distributed random variable Z with mean 0 and standard deviation *sd* to exceed $\pm|m - m'|$ in the respective direction, i.e., *score*(p, p') = ($|Z| \geq |m - m'|$). This score drops exponentially from 1 to 0 with increasing mass difference. As it is always positive, we set the score to $-\infty$ if it falls below 0.05, that is, if the mass difference exceeds $\approx 2 \cdot sd$. A similar approach is taken in ProFound and the tandem MS software SCOPE [19], whereas Mascot uses a constant positive matching score similar to that of Example 1.

Robust incorporation of intensities

In order to incorporate intensities of measured peaks into the scoring, we applied methods from robust statistics successfully used in tandem MS scoring [20]: All peaks in the peak list were ranked according to their absolute intensity. The intensity of the 10% highest abundant peaks were set to 1, the intensity of the 10% lowest abundant peaks to 0. The intensities of the remaining peaks were scaled linearly between 0 and 1. Thus, a very high abundant peak of chemical noise or a small number of wrongly detected, low abundant peaks cannot spoil the interpretation of the whole peak list. Up to this point, we only use intensity values of measured peaks, resulting in an asymmetric scoring scheme. Given an appropriate prediction model [18,21], it would also be possible to incorporate predicted intensities into the scoring. Writing int(p') for a peak's scaled intensity, the matching score is multiplied with (1 + 2 int(p'))/3, yielding a factor of 1/3 for lowest and 1 for highest abundant peaks. Again, the approach is suitable for using any other incorporation of intensity information, such as logarithmic transforms as proposed, e.g., in [22].

Scoring gap peaks

Using peak-wise scoring schemes allows us to explicitly take additional and missing peaks into account.

For additional peaks, a constant penalty c_1 is given. If intensities are used in the scoring, this penalty is again multiplied by the scaled intensity of the peak: *score*(ε, p') = $c_1 \cdot$ int(p'). Very low abundant peaks are then penalized by 0 and thus simply ignored, and very high abundant peaks that are not explained are highly penalized. For missing peaks, the Gaussian score always gives a constant penalty c_2, but as with matching scores, predicted peak intensities could also be used.

Background model and significance of alignment scores

To estimate the significance of a score of a measured spectrum and a sequence of certain length L, we compute a table of mass occurrence probabilities in random weighted strings in a preprocessing step. Using these probabilities, we get a background model for predicted spectra allowing us to estimate the contribution of each measured peak to the overall alignment score under a well-defined null-model without sampling. We proceed as follows: After introducing a formal model of random protein sequences and their digestion, we compute the joint length-mass distribution of cleavage fragments in such random proteins. We then compute the probability that in a random protein of given length, at least one fragment of certain mass m occurs and will thus give rise to a corresponding peak in the predicted spectrum. All these quantities can be computed once in a pre-processing step. Using the mass occurrence probabilities, we estimate the expectation and variance of an alignment score for computing p-values of such the score.

Weighted strings

A *weighted alphabet* is a finite alphabet Σ together with a *weight* or *mass function* μ: $\Sigma \rightarrow \vee_{>0}$, assigning a *mass* to each of its characters. Its domain can be extended to strings $s \in \Sigma^*$ by setting $\mu(s) := \sum_{i=1}^{|s|} \mu(s_i)$. Such strings are called *weighted strings*.

If each character $\sigma \in \Sigma$ occurs with probability (σ), we call an i.i.d. sequence of such characters a *random weighted string*. The parameters of this model, i.e., the character frequencies, can be robustly estimated from a sequence database. As they are the only parameters needed for subsequent significance computations, we can use species-specific models where only small sequence databases are available.

Here, we use the alphabet of amino acids of size 20 together with the molecular mass of the amino acids in Dalton (Da), with 1 Da approximately the weight of a neutron. For the computations, we require the masses to be integers. As measured masses are only known to some precision, we can simply scale the real mass by an appropriate precision factor (0.1 or 0.01 for PMF/MALDI) and denote the resulting integer masses by $\mu^*(\sigma)$. For a precision of 0.1 and character frequencies estimated from

SwissProt, release 48, a sample of the weighted amino acid alphabet is given in Table 3:

This model is readily extendible to capture distributions of masses for each character, such that copies of the same amino acid in a protein may have different masses. This allows to model isotopic mass distributions and different amino acid masses due to post-translational modifications such as phosphorylation or methylation.

Cleavage schemes

Most proteases cleave a peptide right after the occurrence of a specific *cleavage character* in the amino acid sequence, except in the presence of a *prohibition character* directly following the cleavage character. In the case of trypsin, the set of cleavage characters is $\Gamma = \{K, R\}$ and the set of prohibition characters is $\Pi = \{P\}$. Together, Γ and Π form a *cleavage scheme*.

Applying a cleavage scheme on a weighted string results in a *fragmentation* of this string, a set of successive, non-overlapping substrings, the *fragments*.

Example 2 (Fragmentation of a string)

Let $\Sigma := \{A, B, C\}$, be a weighted alphabet with weights $\mu(A) = 1$, $\mu(B) = 2$, $\mu(C) = 3$, let $\Gamma := \{B\}$, $\Pi := \{A\}$ be a cleavage scheme on Σ. Then the string s = ABBACCBACBBB is fragmented into the fragments AB, BACCBACB, B, B of weights $\mu(AB) = 3$, $\mu(BACCBACB) = 17$ and $\mu(B) = 2$.

For the sake of brevity, we concentrate on cleavage schemes without prohibition characters. Then, a fragment is simply a string of non-cleavage characters followed by a cleavage character. Generalizations to arbitrary cleavage schemes and more details on the stochastic models and efficient computation can be found in [10].

Mass occurrence probabilities

Let $f^L[l, m^*]$ denote the probability that the first fragment of a random weighted string of length L has length l and integer mass m^*. The main recurrence is given for the length-mass distribution of the inner part of a fragment, consisting solely of non-cleavage characters:

$$f'[l, m^*] = \sum_{\sigma \notin \Gamma} f'[l-1, m^* - \mu^*(\sigma)] \cdot \mathbb{P}(\sigma),$$

with initial condition $f'[0, 0] = 1$.

The fragment length-mass distribution can be computed by adding the cleavage character to the right and taking care of the finite string length L.

Lemma 1 (Fragment probabilities)

The fragment probability $f^L[l, m^*]$ is given for $l < L$ by

$$f^L[l, m^*] = \sum_{\sigma \notin \Gamma} f'[l-1, m^* - \mu^*(\sigma)] \cdot \mathbb{P}(\sigma)$$

and for the boundary $l = L$ by

$$f^L[L, m^*] = f'[L-1, m^*] + \sum_{\sigma \in \Gamma} f'[L-1, m^* - \mu^*(\sigma)] \cdot \mathbb{P}(\sigma).$$

The probability that a fragment of length l does *not* have mass m is computed as $\bar{f}^L[l, m^*] = u[l] - f^L[l, m^*]$, where $u[l]$ denotes the probability that a fragment has length l; it is a geometric distribution.

Taking the complementary probability $\bar{f}^L[l, m^*]$, we can compute the *mass occurrence probability* $p[L, m^*]$ that at least one fragment of mass m^* occurs in the fragmentation of a random weighted string of length L.

Lemma 2

The occurrence probability $p[L, m^*] = 1 - \bar{p}[L, m^*]$ of mass m^* in a random weighted string of length L is given by $\bar{p}[0, m^*] = 1$ and

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

Table 3: Example weighted amino acids

σ	A (Ala)	C (Cys)	D (Asp)	E (Glu)	...	Y (Tyr)
$\mu(\sigma)$	71.0371	103.0092	115.0269	129.0426	...	163.0633
$\mu^*(\sigma)$	710	1030	1150	1290	...	1631
(σ)	0.0785	0.0154	0.0531	0.0661	...	0.0306

Example weighted alphabet for amino acids. Five amino acids with their average molecular weight in Dalton, derived integer molecular weight with precision 0.1 and relative frequencies estimated from SwissProt, release 48.

Both tables have to be computed up to the largest sequence length L_{max} in the sequence database and up to the largest integer fragment mass $m^*_{\max}$. For PMF using MALDI, $m_{max} \approx$ 3,000 Da and for SwissProt as sequence database, $L_{max} \approx$ 10,000. For a mass precision of one decimal, using doubles, we would need about $30{,}000 \cdot 10{,}000 \cdot 8 \approx 2.24$ GB of main memory for each table. As $\bar{f}^{L}[\cdot,\ \cdot]$ is only needed during computation of $\bar{p}[\cdot,\ \cdot]$, only a very small part of about 3–4 MB is required at any time. To efficiently compute the significance of an alignment score, however, the occurrence probability table $p[\cdot,\ \cdot]$ needs to be kept in memory. Its columns can be computed independently and entries of each column depend smoothly on L (the occurrence probability will not change abruptly if the sequence length grows), it is thus sufficient to store only the first 100 entries of each column completely and then store every 25th row, performing a linear interpolation to get intermediate values. Comparing the exact values in each column to the values computed by the described interpolation scheme, we found the interpolation error to be smaller than 10^{-9} in every case. Note that the interpolation nodes are the exact values, so the interpolation error does not accumulate with growing string length. The mass occurrence probability $p[L, m]$ is given for masses m = 1000.0, 1500.0, 2000.0, 2500.0, 3000.0 Da and a precision of 0.1 Da in Figures 4 and 5, for string length up to 50 and 1000, respectively, showing the continuous behavior of the function for $L > 40$. The "hump" at small string lengths can be explained by the fact that for these lengths, the only possible fragment of mass m is whole the string itself. For greater string length, the corresponding fragment(s) must be "real" fragments, subject to tighter constraints on their combinatorial character composition, e.g. they must have a cleavage character at the end. This "hump" is located around $L \approx m/\mu_{\text{avg}}$, where μ_{avg} denotes the average character mass. For average molecular masses and SwissProt frequencies we have $\mu_{\text{avg}} \approx 111.2$ Da. By further exploiting the fact that $f^L[l, m^*] = 0$ for $l > m^*/\mu^*_{\min}$, where $\mu^*_{\min}$ is the smallest integer character mass in Σ, both $\bar{f}^{L}[l, m^*]$ and $\bar{p}[L, m^*]$ can be computed in time $O(L_{\max} \cdot m^*_{\max})$. We would like to refer the interested reader to [10] for details and proofs on the memory- and time efficient implementation.

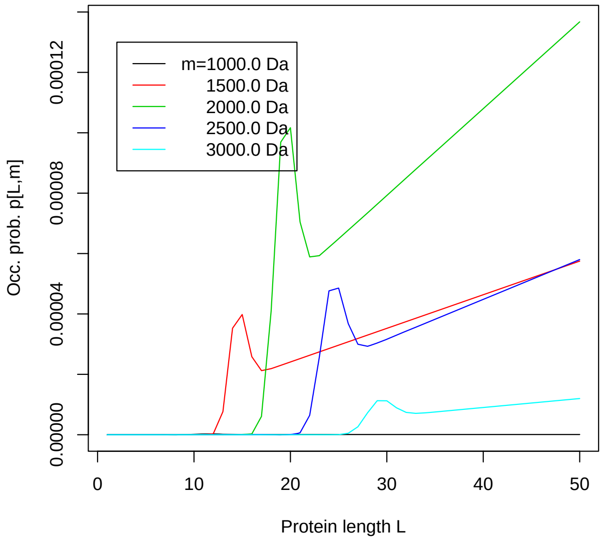

Figure 4
Occurrence probabilities. The mass occurrence probabilities $p[L, m]$ for masses. m = 1000.0, 1500.0, 2000.0, 2500.0, 3000.0 Da and string length L = 1 ... 30. Precision 0.1 Da.

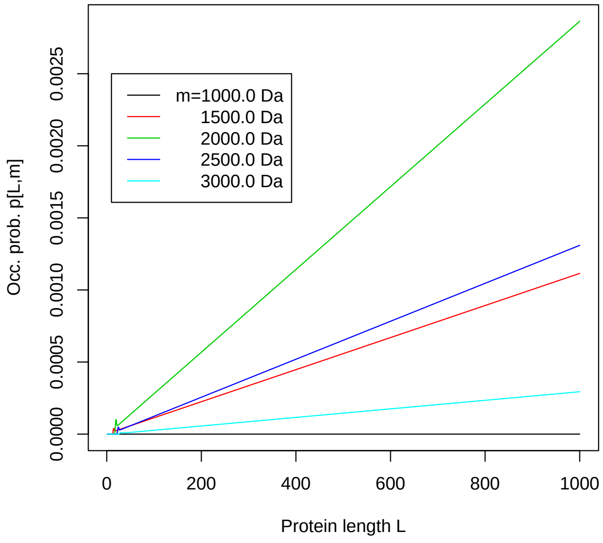

Figure 5
Occurrence probabilities. The mass occurrence probabilities $p[L, m]$ for masses. m = 1000.0, 1500.0, 2000.0, 2500.0, 3000.0 Da and string length L = 1 ... 1000. Precision 0.1 Da.

Alignment score distribution

The alignment score distribution is efficiently and deterministically estimated by adding the contribution of each peak to the overall alignment score. To compute the contribution of each measured peak $p'_j, \in \mathcal{S}_m,$ let $\mathcal{U}_{j\,j}$ denote the *support* of peak $p'_j,$ that is the set of integer masses m^* for which a predicted peak p having integer mass m^* would contribute a positive matching score $score(p, p'_j) > 0$. Let $X_j^{\text{match}}(L)$ be the random variable that contains the sum of the matching scores over all peaks p in any spectrum $\mathcal{S}_p$ generated by a random weighted string of length L that have masses in $\mathcal{U}'_j$. The expectation and variance of this random variable are then given by

$$\mathbb{E}(X_j^{\text{match}}(L)) = \sum_{m^* \in \mathcal{U}'_j} p[L, m^*] \cdot score(p, p'_j)$$

$$\text{Var}(X_j^{\text{match}}(L)) = \sum_{m^* \in \mathcal{U}'_j} p[L, m^*] \cdot score(p, p'_j)^2 - \left(\mathbb{E}(X_j^{\text{match}}(L))\right)^2$$

Similarly, we define random variables $X_j^{\text{add}}(L)$ for the additional scores. Assuming independence of peaks, the overall matching and additional scores are simply the sum of these scores:

$$X^{\text{match}}(L) = \sum_{j=1}^{n'} X_j^{\text{match}}(L) \quad \text{and} \quad X^{\text{add}}(L) = \sum_{j=1}^{n'} X_j^{\text{add}}(L)$$

The missing scores are given for all masses that are not inside the support of any measured peak:

$$X^{\text{miss}} = \sum_{m^* \notin \bigcup \mathcal{U}'_j} X_{m^*}^{\text{miss}}$$

We omit the details for additional and missing peaks and again refer the interested reader to [9]. The alignment score $score(\mathcal{S}_p, \mathcal{S}_m)$ for a measured spectrum $\mathcal{S}_m$ and a random predicted spectrum $\mathcal{S}_p$ generated by a random weighted string of length L is finally given by

$$score(\mathcal{S}_p, \mathcal{S}_m) = X^{\text{match}}(L) + X^{\text{add}}(L) + X^{\text{miss}}.$$

As $score(\mathcal{S}_p, \mathcal{S}_m)$ is the sum of nearly independent random variables, we can expect it to have a normal distribution for reasonable scoring schemes and peak lists. This distribution is completely determined by its expectation and variance.

Note that the alignment algorithm computes an optimal one-to-one peak matching score, whereas the estimation procedure corresponds to a many-to-one matching of peaks. As shown below, neither the peak independence assumptions nor the one-to-one peak matching are a problem in practice, as violations of either assumption do not contribute enough to alter the score distribution noticably.

We tested the two assumptions using two different parameter sets A and B for the Gaussian score given in Table 4, and computing the alignment scores of 10,000 random amino acid sequences of length 250 and a randomly chosen measured spectrum from our dataset. We found the estimated distributions in good agreement with their empirical counterparts, as shown in Figures 6 and 7.

Using significance as score

As the alignment score is an additive score, its value and distribution is dependent on the number of peaks in the measured and predicted spectra. This makes it difficult to compare alignment scores for different measured spectra and sequence lengths.

To avoid these problems, we will not use the alignment score itself, but its significance to rank the candidate sequences, a method previously shown to be effective for tandem MS data [20]. For each pair of measured and predicted spectra, the alignment score is computed, its distribution is estimated using the method described, and the p-value – the probability that a random sequence of the same length as the aligned sequence gives an alignment score at least as good as the computed one – is computed from this distribution. We then take – $\log_{10}$(p-val.) as the *significance score* to rank the candidates. In the evaluation, we always used the significance score unless explicitly stated otherwise.

Table 4: Parameter sets for evaluation

Parameter	std. dev. *sd*	missing score	additional score	intensity used
A	0.8	-0.1	-0.1	No
B	0.8	-0.4	-0.3	Yes

The two parameter sets A and B used for the evaluation. Shown are the standard deviation for the Gaussian mass deviance distribution, the missing and additional scores and whether relative peak intensities are used.

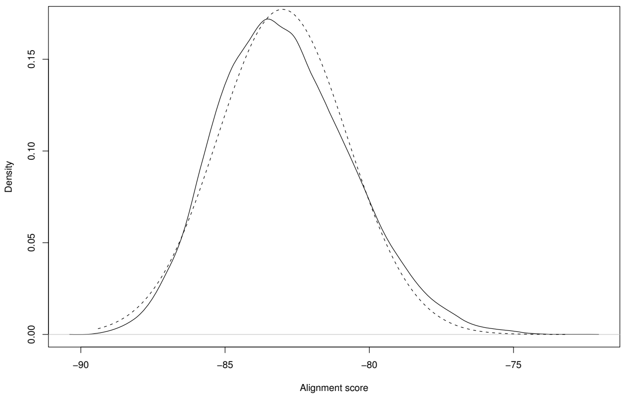

Figure 6

Alignment score distribution. **A**: Solid line: Densities of empirical alignment score distribution using 10,000 randomly generated protein sequences of length 250 with SwissProt amino acid frequencies. Dashed line: Density of approximating normal distribution with parameters computed as described in the text. Both alignments for one measured spectrum and SAMPI score with parameter set A.

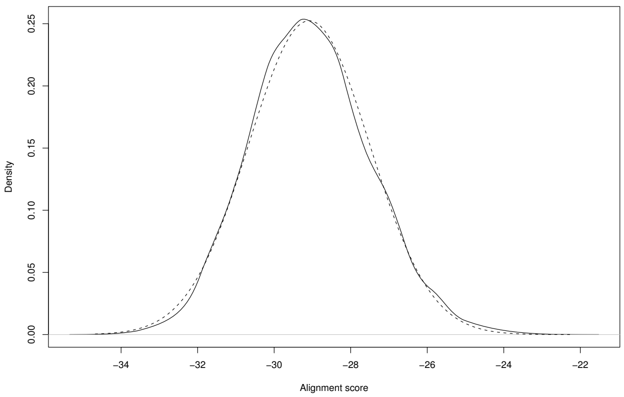

Figure 7

Alignment score distribution. **B**: Solid line: Densities of empirical alignment score distribution using 10,000 randomly generated protein sequences of length 250 with SwissProt amino acid frequencies. Dashed line: Density of approximating normal distribution with parameters computed as described in the text. Both alignments for one measured spectrum and SAMPI score with parameter set B.

Authors' contributions

HMK and SB developed the alignment formulation and score statistics. HMK performed the SAMPI experiments and evaluation and wrote the manuscript. AW provided the mass spectrometry data and performed the Mascot experiments. All authors proof-read and approved the manuscript.

Acknowledgements

Hans-Michael Kaltenbach is supported by the "International NRW Graduate School in Bioinformatics and Genome Research." Implementation of the presented method was provided by Tobias Marschall, Marcel Martin, Matthias Steinrücken, and Henner Sudek. The authors want to thank Sven Rahmann for helpful discussions and Martina Mahne for providing the original mass spectrometry data.

References

1. Aebersold R, Mann M: **Mass spectrometry-based proteomics.** *Nature* 2003:198-207.
2. Bairoch A, Boeckmann B: **The SWISS-PROT protein sequence data bank.** *Nucleic Acids Res* 1992, **20:**2019-2022.
3. Henzel WJ, Watanabe C, Stults JT: **Protein Identification: The Origins of Peptide Mass Fingerprints.** *J Am Soc Mass Spectrometry* 2003, **14:**931-942.
4. Sadygov RG, Cociorva D, Yates JR: **Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book.** *Nat Methods* 2004, **1(3**195-202 [http://dx.doi.org/10.1038/nmeth725].
5. Karas M, Hillenkamp F: **Laser desorption ionization of proteins with molecular masses exceeding 10,000 Daltons.** *Anal Chem* 1988, **60:**2299-2301.
6. Perkins D, Pappin D, Creasy D, Cottrell J: **Probability-based protein identification by searching sequence databases using mass spectrometry data.** *Electrophoresis* 1999, **20:**3551-3567.
7. Zhang W, Chait BT: **ProFound: an expert system for protein identification using mass spectrometric peptide mapping information.** *Anal Chem* 2000, **72(11):**2482-2489.
8. Chamrad DC, Körting G, Stühler K, Meyer HE, Klose J, Blüggel M: **Evaluation of algorithms for protein identification from sequence databases using mass spectrometry data.** *Proteomics* 2004, **4:**619-628.
9. Böcker S, Kaltenbach HM: **Mass Spectra Alignments and Their Significance.** In *Combinatorial Pattern Matching Volume 3537*. Edited by: Apostolico A, Crochemore M, Park K. Springer; 2005:429-441.
10. Kaltenbach HM, Sudek H, Böcker S, Rahmann S: **Statistics of cleavage fragments in random weighted strings.** *Tech. Rep. TR-2005-06, Technische Fakultät der Universität Bielefeld, Abteilung Informationstechnik* 2005 [http://bieson.ub.uni-bielefeld.de/volltexte/2006/900/].
11. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.
12. Havilio M, Haddad Y, Smilansky Z: **Intensity-based statistical scorer for tandem mass spectrometry.** *Anal Chem* 2003, **75:**435-444.
13. Wilke A, Rückert C, Bartels D, Dondrup M, Goesmann A, Hüser AT, Kespohl S, Linke B, Mahne M, McHardy AC, Pühler A, Meyer F: **Bioinformatics support for high-throughput proteomics.** *J Biotechnol* 2003, **106(2–3):**147-56.
14. Gusfield D: *Algorithms on Strings. Trees, and Sequences* Cambridge University Press; 1997.
15. Huang X, Waterman MS: **Dynamic programming algorithms for restriction map comparison.** *Comput Appl Biosci* 1992, **8(5):**511-520.
16. Wenk C: **Applying an Edit Distance to the Matching of Tree Ring Sequences in Dendrochronology.** *Proceedings of Combinatorial Pattern Matching (CPM99)* 1999, **1645:**223-242.
17. Bylund D, Danielsson R, Malmquist G, Markides KE: **Chromatographic alignment by warping and dynamic programming as a pre-processing tool for PARAFAC modelling of liquid chromatography-mass spectrometry data.** *Journal of Chromatography A* 2002, **961:**237-244.

18. Gay S, Binz PA, Hochstrasser DF, Appel RD: **Peptide mass fingerprinting peak intensity prediction: Extracting knowledge from spectra.** *Proteomics* 2002, **2:**1374-1391.
19. Bafna V, Edwards N: **SCOPE: A probabilistic model for scoring tandem mass spectra against a peptide database.** *Bioinformatics* 2001, **17:**S13-S21.
20. Wan Y, Yang A, Chen T: **PepHMM: A Hidden Markov Model Based Scoring Function for Mass Spectrometry Database Search.** In *Proc of RECOMB 2005 Volume 3500*. Springer; 2005:342-356.
21. Schütz F, Kapp EA, Simpson RJ, Speed TP: **Deriving statistical models for predicting peptide tandem MS product ion intensities.** *Biochem Soc Trans* 2003, **31(Pt 6):**1479-1483.
22. Wolski WE, Lalowski M, Martus P, Herwig R, Giavalisco P, Gobom J, Sickmann A, Lehrach H, Reinert K: **Transformation and other factors of the peptide mass spectrometry pairwise peak-list comparison process.** *BMC Bioinformatics* 2005, **6:**285.