Efficient Interpretation of Tandem Mass Tags in Top-Down Proteomics

Anna Katharina Hildebrandt[1], Ernst Althaus[2], Hans-Peter Lenhof[1], Chien-Wen Hung[3], Andreas Tholey[3], and Andreas Hildebrandt[2]

1 **Center for Bioinformatics, Saarland University, 66041 Saarbrücken, Germany**
{anna.hildebrandt, lenhof} @ bioinf.uni-sb.de
2 **Johannes-Gutenberg-University Mainz, 55128 Mainz, Germany**
{andreas.hildebrandt, ernst.althaus} @ uni-mainz.de
3 **Institut für Experimentelle Medizin, Kiel University, 24105 Kiel, Germany**
{cw.hung, a.tholey} @ iem.uni-kiel.de

Abstract

Mass spectrometry is the major analytical tool for the identification and quantification of proteins in biological samples. In so-called top-down proteomics, separation and mass spectrometric analysis is performed at the level of intact proteins, without preparatory digestion steps. It has been shown that the tandem mass tag (TMT) labeling technology, which is often used for quantification based on digested proteins (bottom-up studies), can be applied in top-down proteomics as well. This, however, leads to a complex interpretation problem, where we need to annotate measured peaks with their respective generating protein, the number of charges, and the a priori unknown number of TMT-groups attached to this protein. In this work, we give an algorithm for the efficient enumeration of all valid annotations that fulfill available experimental constraints. Applying the algorithm to real-world data, we show that the annotation problem can indeed be efficiently solved. However, our experiments also demonstrate that reliable annotation in complex mixtures requires at least partial sequence information and high mass accuracy and resolution to go beyond the proof-of-concept stage.

1998 ACM Subject Classification J.3 Biology and genetics, G.1.6 Optimization

Keywords and phrases Mass spectrometry, TMT labeling, Top-down Proteomics

Digital Object Identifier 10.4230/OASIcs.GCB.2013.56

1 Introduction

The two major goals of proteomics are to identify and quantify proteins present in a given sample. Today, the most important analytical technique for this purpose is mass spectrometry (MS). Typical protein mixtures are highly complex: proteomes contain hundreds to several hundreds of thousands of proteins and protein forms. Often, many components of the sample will have similar masses, leading to overlapping signals in the spectrum that are hard to disentangle. Hence, the proteins usually have to be separated prior to MS with respect to a property that is not strongly correlated with the mass; a powerful technique for this purpose is liquid chromatography (LC).

The separated samples are then injected into a mass spectrometer, leading to a series of mass spectrometric runs, each applied on the sample content eluting from the chromatographic column at a specific retention time (RT) interval. In the mass spectrometer, the molecules are then ionized by the attachment of z protons (simultaneously delivering z positive charges), and accelerated in an electric field. Since the reaction of the peptide to the field depends on

German Conference on Bioinformatics 2013 (GCB'13).
Editors: T. Beißbarth, M. Kollmar, A. Leha, B. Morgenstern, A.-K. Schultz, S. Waack, E. Wingender; pp. 56–67

OpenAccess Series in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

the ratio $\frac{m}{z}$ of peptide mass over its acquired charge, this quantity can now be measured. In realistic spectra, the informative parts of the mass spectrum (the 'signal' content) have to be identified and separated from parasitics, such as high-frequency noise or low-frequency baseline terms [5]. The resulting parts of the signal that are believed to arise from a molecule of interest are known as *peaks*. MS signals are usually formed by groups of peaks, representing the sum of all isotopes contained in the molecule. Consequently, if the spectrometer records a peak for a molecule with mass m_i at $\frac{m_i}{z}$, we can typically expect to find a peak also at $\frac{m_i+m_p}{z}$, where m_p is the mass of a single proton. If the spectral resolution allows to separate and identify at least two successive isotopic peaks, the molecular charge can be inferred.

With this information, we can now try to identify the molecular content of the sample. In proteome analysis, we are typically given a database with the amino acid sequences of potentially occurring proteins. In the general case, this database might be comprised of all proteins contained in, e.g., Uniprot [2] for the species of interest. The masses found in the experiment are then used as a query against the database. But unfortunately, the molecular mass alone is often not sufficiently characteristic for the molecule. Trivially, all sequence permutations of a given protein lead to the same mass and, hence, cannot be distinguished from this information alone. Even mutations of the sequence often lead to mass differences that are too small to be recognizable.

In *tandem mass spectrometry*, or *MS/MS*, this problem is solved by fragmentation of those parts of the sample that were identified to be potentially relevant. Since proteins preferably break at well-defined positions along their backbone, a large enough sample of the fragmentation space (induced, e.g., through molecular collisions) will lead to pairs of corresponding masses from which at least partial sequence information can be derived. Since this information characterizes the molecule much better MS/MS is typically required today for reliable identification.

For molecules as large as proteins, many steps of this procedure become very challenging. For instance, the MS/MS spectra of proteins are much harder to interpret than those of smaller molecules, and their isotopic patters are much more complex. Thus, proteins are often first digested into peptides with the help of specific proteases. The query database is then virtually digested: for a given protein sequence, the resulting peptides can be easily inferred from the protein's primary sequence, since the restriction enzymes cut the sequence at specific cleavage sites[1]. From the identification of the peptides found in the mass spectrum, we try to infer the proteins that contained those peptides. To this end, different scoring schemes [3, 10] based on different statistical models can be used to generate p-values for the occurrence of the proteins in the database.

Such a setup, with its digestion of proteins into peptides, which are then identified and used as evidence for their containing proteins, is known as *shotgun-* or *bottom-up* proteomics. The major advantage of this technology is that peptides are much simpler to separate by LC and can be measured with higher mass accuracy and sensitivity in mass spectrometry. The MS/MS spectra of peptides are easier to interpret, even though in many cases a large percentage of them cannot be annotated successfully. Thus, even though bottom-up proteomics is a very sensitive method, many of the peptides are missed in practice. Also, some of the digestion peptides for a given protein might not ionize sufficiently well to allow their detection, or they might be too small or too large for the given experimental setup. This often leads to non-optimal coverage of the protein sequence by peptides in the

[1] Since the protease might miss potential cuts, it is customary to generate all peptide combinations up to a fixed number of missed cleavage sites.

digest: even though the protein has been identified through several of its digestion peptides, large parts of its sequence might not be represented in the results. In summary, the increased complexity of the samples – out of a single protein, multiple peptides are generated – imposes challenges even though each individual component is more easily identified.

These drawbacks are avoided by so-called *top-down* proteomics studies, where no digestion of the protein into peptides is performed. Instead, the sample is separated via LC at the level of intact proteins, at which also the MS experiment is performed. Hence, the information belonging to a single protein is not distributed over many peptides, which allows direct distinction of protein isoforms or of post-translationally modified forms from their non-modified counterparts. However, the separation of intact proteins is not as straightforward as that of peptides, and the detection limits of proteins in MS are strongly elevated with increasing protein size. The situation is complicated further, as from the MS/MS spectra of intact proteins, only limited information about N- and C-terminal parts of the protein sequence can be derived, hampering unambiguous identification. Nevertheless, the clear advantages for downstream analysis due to an automatically full sequence coverage of the proteins make top-down proteomics an increasingly popular alternative to bottom-up studies.

An even greater challenge than the identification of proteins is their accurate quantification. Several strategies for relative as well as for absolute quantification have been proposed [1]. In addition to label-free approaches, methods for quantification in MS mode using stable isotope labeling quantification have been developed, where the molecules of interest are modified with chemical groups that allow for an accurate quantification. Another approach is the use of isobaric labeling strategies, where the samples are labeled with reagents which consist of three major groups: (i) a reactive group that allows covalent attachment of the reagent to the peptide, in particular the N-termini and epsilon-amino groups of Lysine residues; (ii) a reporter group and (iii) a balancer group. These reagents can now be formed in four or eight (iTRAQ [4]) or six (TMT [13]) different flavors: for TMT, for instance, a reporter group in the first flavor has a mass of 126 Da, the corresponding balancer of 103 Da, yielding a total mass of the label of 229 Da. The reporter of the second flavor has a mass of 127 Da, the balancer of 102 Da, again yielding the same total mass of 229 Da. As the different flavors share the same molecular properties and only differ in the isotope composition, they also appear at the same retention times. Hence, after labeling of different biological samples with these reagents (one flavor per sample), the labeled samples can be combined, treated by LC, and analyzed by MS. In MS mode, equivalent peptides from different biological samples will have the same m/z-values, as the reagents were isobaric. But upon fragmentation of the peptides in MS/MS experiments, the reporter groups are liberated, yielding signals of the corresponding reporter ions at 126, 127, ..., 131 Da. The intensities of these reporter ion signals deliver a direct readout of the relative quantities of the peptides in the different biological samples. Isobaric labeling strategies have been originally developed for quantification studies in bottom-up approaches, i.e., at the level of peptides. But a recent pilot study [6] has shown that tandem mass tag labelling can be applied in a top-down setting as well. At this stage, the method is still restricted to the quantification and simultaneous MS/MS-based identification of relatively small proteins up to ≈ 35 kDa. A severe bottleneck is the interpretation of MS- and MS/MS-spectra of such isobarically labeled intact proteins, which will be the focus of this work.

Three different effects render this a challenging task: (i) The degree of labeling may differ and is unknown for an a-priori unknown protein. This is caused by incomplete labeling or unwanted non-specific labeling of residues in proteins. Consequently, a theoretical protein with 15 Lysine residues can lead to a mixture containing protein species with 12 to about

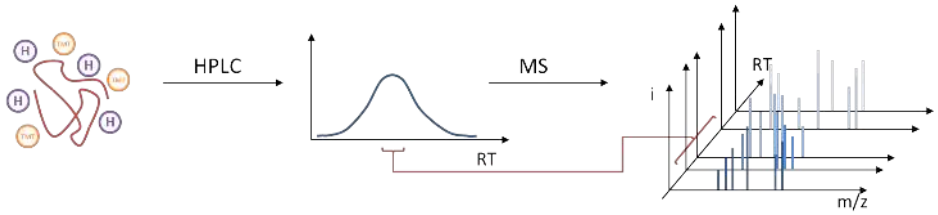

Figure 1 The experimental setup of top-down tandem mass tag proteomics [6]

18 isobaric labels attached. (ii) Each of the proteins features a complex isotopic pattern as outlined above. (iii) In the ionization process (electrospray) used for the analyis of intact proteins, species with different charge states are formed by attachment of different numbers of protons. Thus, a single theoretical protein with a mass of 15 kDa may have 10, 11, ..., 20 attached protons, leading to peak groups at $\frac{15}{10}$ kDa, $\frac{15}{11}$ kDa, ..., $\frac{15}{20}$ kDa. This number of attached protons cannot be predicted, in particular not for unknown proteins in unknown proteomes. But it can be deduced from the mass differences of the isotopes in the peak group according to the relation given above.

These factors lead to a difficult interpretation problem, where we want to analyze for each peak which proteins could have generated it, and compute the corresponding number of charges[2] and TMT marker groups. In this work, we will present an efficient algorithm for this annotation task and will apply it to real-world experimental data. Using our algorithm, we will further demonstrate that the information contained in the experiment is insufficiently specific, resulting in false-positive annotations that match the given masses. We will then discuss how to integrate further experimental insight into the algorithm at moderate computational cost that can weed out many false positives.

2 Methods

2.1 Experimental setup

For generation of MS and MS/MS datasets, a mixture of six known model proteins was labeled with the TMT-6-plex reagent. It has to be noted that several of the six model proteins contained impurities, thus finally ten different proteins were present in the test mixture (see Tab. 1). The proteins were separated via ion paring reversed phase chromatography using monolithic columns and analysed in a Thermo Orbitrap Velos mass spectrometer equipped with ETD. The mass spectrometer was operated in the data-dependent mode to switch automatically between Full-MS (scan 1), HCD-MS^2 (scan 2), and ETD-MS^2 (scan 3). After a Full-MS scan acquired in the ion trap MS, the most abundant protein ion (top 1) was selected for an HCD-MS^2 scan and an ETD-MS^2 scan. Full details of the experimental procedure were described in [6]. A schematic sketch of the approach is shown in Fig. 1.

[2] Please note that, in principle, other ionization types than addition of protons can occur, and can indeed be handled by our method. For reasons of simplicity, these will not be considered in the current manuscript.

Table 1 The ten protein mix and the results of the manual annotation [6]. Note that in the manual annotation, all Lysine residues were assumed to carry a TMT group. The mix consists of six known model proteins labeled with the TMT-6-plex reagent and four impurities.
* : C-terminal seq. of ovalbumin - ASVSEEFRADHPFLFCIKHIATNAVLFFGRCVSP
#: undefinable in manual annotation due to the poor MS quality derived from post-translational modifications (e.g. phosphorylation, glycosylation etc)

Protein Name	Theo. MW (kDa)	No. Of Cys/Lys	Theo. MW with TMT (kDa)	Observed m/z	Charge State	Calc. MW (Da)
Cytochrome C *(Equine)*	12.3	2C/19K	16.8	1045.7 1115.2 1194.8	16 15 14	16714.3 16713.2 16713.0
Myoglobin *(Equine)*	16.9	0C/19K	21.5	1436.6 1539.1 1657.4	15 14 13	21533.9 21533.3 21533.1
Carbonic Anhydrase *(Bovine)*	29.1	0C/18K	33.2	949.7 977.7 1007.2	35 34 33	33205.6 33208.9 33205.3
Carb. Anhydrase Impurity 1 **Ubiquitin** **partial seq.**	8.5	0C/7K	10.2	1017.9 1130.8 1272.0	10 9 8	10168.9 10168.1 10168.2
Carb. Anhydrase Impurity 2 **Superoxide Dismutase**	15.5	3C/10K	17.9	1278.4 1376.7 1491.2	14 13 12	17882.8 17883.7 17882.2
Ovalbumin *(Gallus)*	42.8	6C/20K	47.7	UD#		
Ovalbumin Impurity 1 **Ovomucoid**	20.1	18C/13K	24.3	UD#		
Ovalbumin Impurity 2 **C-terminal ovalbumin***	3.8	2C/1K	4.3	877.5 1096.5 1461.9	5 4 3	4382.5 4381.9 4382.7
BSA *(Bovine)*	66.6	34C/60K	82.3	UD#		
Apo-transferrin *(Bovine)*	77.7	38C/64K	94.5	UD#		

2.2 Formal problem formulation: The TMT annotation problem

In this section, we will introduce the formal definition of the annotation problem posed by top-down TMT labelling. Informally, we want to query a database of known protein masses (e.g., the whole proteome of the organisms contained in the sample) against the peaks detected in the experiment. To this end, we want to decide for every protein in the database and for every peak, whether this protein could have led to the peak's observed mass-over-charge ratio through a feasible combination of base protein mass, TMT attachments, and charges (protons). To formally formulate the problem, we first need a few definitions.

Let m_T denote the mass of the TMT marker group ($m_T \approx 229.162932$ Da), and m_p the mass of a single proton. By $DB := \{m_i | i = 1, \dots n_{DB}\}$, we denote the database we want to query, where m_i is the monoisotopic mass of the i-th protein in DB. We assume that the spectrum has been pre-processed to yield a set of mass spectrometric peaks $\mathcal{S} := \{p_j | j = 1, \dots n_{\mathcal{S}}\}$.

Let us further assume that one of the populations in the sample was given by the i-th protein, to which β_i TMT-groups and α_i excess protons have been attached, with $\beta_i, \alpha_i \in \mathbb{N}^+$. This protein will have a charge of $z = \alpha_i$, measured in units of elementary charge, and a total mass-over-charge ratio of

$$m_{z,i}\left(\alpha_i, \beta_i\right) := \frac{m_i + \alpha_i m_p + \beta_i m_T}{\alpha_i}$$

This relation between protein, ionization state, and TMT assignment is not unique: one protein species may acquire different ionization states as well as different numbers of attached TMT groups. However, in practice, not all values of α_i and β_i are possible: the amount of charges and of TMT groups that a given protein can acquire falls within limited ranges, i.e.,

$$\alpha_i \in \{\alpha_i^{min}, \ldots \alpha_i^{max}\} \quad \text{and} \quad \beta_i \in \{\beta_i^{min}, \ldots \beta_i^{max}\}$$

In the following we describe how to efficiently reduce the parameter search space. Since the TMT markers attach to Lysine-residues, it is natural to choose $\beta_i^{min}, \beta_i^{max}$ accordingly, and hence limit the number of TMT-attachments to a 2x window, i.e.:

$$\beta_i \in \{\max(0, \#\text{LYS} - x), \ldots \#\text{LYS} + x\} \text{ for } x \in \mathbb{N}^+$$

We now want to annotate all *measured* peaks $p_j \in \mathcal{S}$ with all *predicted* peaks $\hat{p_i}$ due to feasible protein/TMT/proton combinations within a given accuracy threshold. As the accuracy of mass spectra is typically dependent on the mass-over-charge ratio, it is customary to use relative measures of error. Thus, for every measured peak p_j we want to determine all feasible predicted peaks $\hat{p_i}$ with

$$\frac{|\hat{p_i} - p_j|}{p_j} \leq \epsilon$$

To solve this problem, we will compute for every protein m_i in DB and for every peak p_j all feasible values of $\alpha_{i,j}$ and $\beta_{i,j}$, such that the assumption of protein i with $\alpha_{i,j}$ attached protons and $\beta_{i,j}$ attached TMT markers explains peak p_j within the given relative accuracy threshold ϵ, which gives the following combinatorial problem:

$$\begin{array}{l}
\forall i \in \{1, \ldots, n_{DB}\}: \\
\quad \forall j \in \{1, \ldots, n_S\}: \\
\quad\quad \text{find } \alpha_{i,j} \in \{\alpha_{i,j}^{min}, \ldots, \alpha_{i,j}^{max}\}, \\
\quad\quad\quad \beta_{i,j} \in \{\max(0, \#\text{LYS} - x), \ldots \#\text{LYS} + x\}: \quad |m_{z,i}(\alpha_{i,j}, \beta_{i,j}) - p_j| \quad \leq \quad \epsilon \cdot p_j
\end{array}$$

Obviously, not all combinations of $\alpha_{i,j}, \beta_{i,j}$ will lead to a valid annotation. In the following we will determine reasonable boundaries for $\alpha_{i,j}$, given $\beta_{i,j}$:

$$\begin{array}{rrcl}
 & |m_{z,i}(\alpha_{i,j}, \beta_{i,j}) - p_j| & \leq & \epsilon \cdot p_j \\
\Leftrightarrow & \left|\frac{m_i + \alpha_{i,j} m_p + \beta_{i,j} m_T}{\alpha_{i,j}} - p_j\right| & \leq & \epsilon \cdot p_j \\
\overset{\alpha_{i,j} > 0}{\Longleftrightarrow} & |m_i + \alpha_{i,j} m_p + \beta_{i,j} m_T - \alpha_{i,j} p_j| & \leq & \alpha_{i,j} \cdot \epsilon \cdot p_j
\end{array}$$

In practice, the allowed window $2x$ for the parameter $\beta_{i,j}$ is quite small (in our experiments, we used $x = 3$). We can thus easily test all allowed values for $\beta_{i,j}$. For any such fixed but

arbitrary $\beta_{i,j}$, we find

$$\begin{array}{lrcrl} & |m_i + \beta_{i,j} m_T + \alpha_{i,j}(m_p - p_j)| & \leq & \alpha_{i,j} \cdot \epsilon \cdot p_j & \\ \Leftrightarrow & (m_i + \beta_{i,j} m_T + \alpha_{i,j}(m_p - p_j)) & \leq & \alpha_{i,j} \cdot \epsilon \cdot p_j & \wedge \\ & (m_i + \beta_{i,j} m_T + \alpha_{i,j}(m_p - p_j)) & \geq & -\alpha_{i,j} \cdot \epsilon \cdot p_j & \\ \Leftrightarrow & \alpha_{i,j}(m_p - p_j - \epsilon \cdot p_j) & \leq & -(m_i + \beta_{i,j} m_T) & \wedge \\ & \alpha_{i,j}(m_p - p_j + \epsilon \cdot p_j) & \geq & -(m_i + \beta_{i,j} m_T) & \end{array}$$

Remembering that m_p denotes the mass of a single proton, and p_j a mass-to-charge ratio in a feasible range for proteins, we see that $m_p \ll p_j$. In addition, the accuracy threshold ϵ is small in practice – on the order of tens or hundreds of parts per million (ppm) – so that we also find $\epsilon \cdot p_j \ll p_j$. Indeed, we can safely assume that $m_p + \epsilon \cdot p_j \ll p_j$ and, hence,

$$(m_p - p_j - \epsilon \cdot p_j) < 0 \quad \wedge \quad (m_p - p_j + \epsilon \cdot p_j) < 0$$

We thus find:

$$\alpha_{i,j} \geq -\frac{m_i + \beta_{i,j} m_T}{m_p - p_j - \epsilon \cdot p_j} \qquad \wedge \qquad \alpha_{i,j} < -\frac{m_i + \beta_{i,j} m_T}{m_p - p_j + \epsilon \cdot p_j}$$

We can thus restrict $\alpha_{i,j}$ as a function of the fixed but arbitrary $\beta_{i,j}$. For simplicity of notation, we introduce $a_{i,j} := m_i + \beta_{i,j} m_T > 0$ and $b_j := p_j - m_p > 0$, which yields

$$\begin{array}{lrcrcrl} & \alpha_{i,j} & \geq & \frac{-a_{i,j}}{-b_j - \epsilon \cdot p_j} & = & \frac{a_{i,j}}{b_j + \epsilon \cdot p_j} & > \quad 0 \\ \wedge & \alpha_{i,j} & < & \frac{-a_{i,j}}{\epsilon \cdot p_j - b_j} & = & \frac{a_{i,j}}{b_j - \epsilon \cdot p_j} & > \quad 0 \end{array}$$

so that we finally obtain

$$\alpha_{i,j}^{min} := \left\lceil \frac{a_{i,j}}{b_j + \epsilon \cdot p_j} \right\rceil, \qquad \alpha_{i,j}^{max} := \left\lfloor \frac{a_{i,j}}{b_j - \epsilon \cdot p_j} \right\rfloor$$

Thus, we only have to consider $\alpha_{i,j} \in \{\alpha_{i,j}^{min}, \ldots, \alpha_{i,j}^{max}\}$. Each of these values will lead to a valid explanation of the peak, i.e., the triple $< m_i, \alpha_{i,j}, \beta_{i,j} >$ yields an m/z-value that deviates from p_j by less than ϵ. We can thus trivially enumerate all valid annotations.

2.3 Results of the procedure

We implemented the scheme described above in OpenMS [9]. To test the correctness, efficiency, and utility of our approach, we applied our implementation to the experimental data set used in [6]. The parameters used in our study were chosen to conform with experience gathered by our experimental partners. For the limits on the number of attachable TMT groups, we used

$$\beta_{i,j}^{min} := (\#LYS + 1) - 2, \quad \beta_{i,j}^{max} := (\#LYS + 1) + 3$$

where the +1 accounts for attachment at the N-terminus. The accuracy threshold was varied to study its influence on the number of generated solutions. Typical values are in the order of tens to hundreds parts per million, i.e., $\epsilon \approx 10^{-5}$ to 10^{-4}. While the minimal and maximal number of charges that can explain a given peak have been computed as a function of $\beta_{i,j}$ above, we also enforce global limits on them. Obviously, we require $\alpha_{i,j} > 0$. In accordance with experimental insight, we also introduce an upper boundary: $\alpha_{i,j} < 40$.

Table 2 Results on the 10-protein mix from [6]. ϵ is given in ppm.

ϵ	annotated peaks	valid annotations	identified proteins
10	510	546	8
20	992	1092	8
100	5090	7665	8
300	12614	30408	8

In [6], the proteins were first subjected to certain chemical modifications, as is commonly the case in proteomics. To account for these modifications in our study, we performed a virtual carboxyaminomethylation. This modification changes the mass of Cysteine-residues by $\delta_m^{\text{carb}} \approx +57.0214$ Da. We assume this modification to be fully effective ("fixed"), and hence arrive at a mass difference of $\#\text{CYS} \cdot \delta_m^{\text{carb}}$ for every protein in the reference database. In addition, every initiator Methionine might be removed, with an acetylation of the new N-terminus, yielding a mass difference of $\delta_m^{\text{acet}} \approx -89.0299$ Da. To account for this variable modification, we add a modified and an unmodified variant for each protein to the reference database.

With these preparations, we first attempted to recreate the results described in [6]. To this end, the reference database was set to contain the 10 proteins known to be present in the sample (for details, see [6]). The spectrum was processed using the OpenMS Wavelet-based peak picker [8, 7, 12], leading to 16,042 peaks. The runtime of our algorithm did not change significantly with varying values of ϵ, and was ≈ 0.55 seconds in each case. Tab. 2 describes the results: 8 out of 10 proteins were consistently found, but many more valid annotations were found than previously expected.

The method as described above uses total mass as a descriptor for a protein, which is known to be insufficiently specific in the general case. Still, the large number of valid annotations came as a surprise for two reasons: first, since the query database is small, only very few proteins were expected to fit a given peak (for larger data bases, partial sequence information derived from MS/MS experiments can help in filtering false positive protein identifications). Second, as a result of the time-consuming manual annotation process used previously, it was suspected that only a small number of TMT/charge-combinations of a given protein would explain any given peak, if it could be explained at all.

In practice, though, the mass accuracy achieved in the ion trap MS is insufficient to rule out many of the possible explanations. For low mass accuracy, several of the peaks of our spectrum could be reasonably explained by multiple variants. For the determination of intact protein masses, the ion trap was used [6]; here the mass accuracies achievable for intact proteins are at best greater than 10 ppm for very small proteins and can reach the Da range for larger ones, yielding very large ppm values[3]. Accordingly, we also performed our experiments with large mass deviations. However, for deviations as large as $\epsilon = 100$ ppm, we could often not even distinguish between acetylated and non-acetylated versions of the protein. We thus decided to adapt the algorithm for improved specificity.

[3] These large mass deviations also prevent to distinguish individual peaks in the isotopic pattern and, hence, a simple determination of the charge. This also prevents application of most feature detection procedures used in bottom-up proteomics.

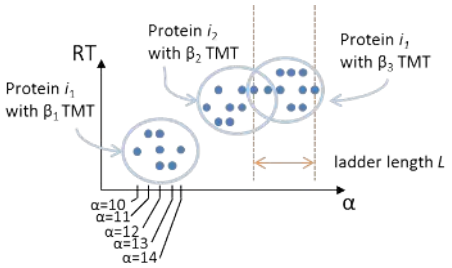

Figure 2 Potential charge ladders of annotated neighbouring peaks (blue nodes). These can be used for filtering significant hits.

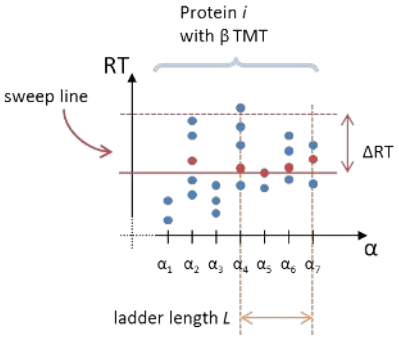

Figure 3 Sweep line (red nodes) traversing all charge variants (i, α_k, β) of protein i having β TMT-assignments.

2.4 Refined problem formulation

As by design every annotation computed by our algorithm is valid, i.e., falls within a chemically reasonable range of TMT- and proton-number, ruling out explanations will require either improved mass accuracy to allow for reducing the threshold ϵ, or the use of additional information. One restriction that can be obtained without any further experimental effort stems from what we call the *charge-ladder assumption*.

Assume that the sample contains a species of protein i with $\alpha_{i,j}$ attached charges and $\beta_{i,j}$ TMT markers. According to experimental experience, it is then very likely that the sample contains the same protein with the same number of TMT markers, but with a charge that is smaller or greater by one. We thus call two annotations $< m_i, \alpha_{i,j}, \beta_{i,j} >$ and $< m_i, \alpha_{i,k}, \beta_{i,k} >$ *neighbours*, iff $|\alpha_{i,j} - \alpha_{i,k}| = 1$. Consequently, we should expect that if we can explain a peak with a given annotation, we can also explain other peaks in the spectrum by its neighbours. Indeed, a manual annotator would reject an explanation if it would not be supported by a gap-less chain of neighbouring annotations. Detection of such chains, however, is complicated by the fact that the corresponding peaks may not necessarily co-elute and, hence, occur at different retention times. But since the physico-chemical change of the neighbouring protein species is small, we expect the neighbouring peaks to be located within a certain retention time interval.

To assign an annotation with a high probability of occurrence, we thus now demand the existence of a *charge ladder* (c.f. Fig. 2) of a minimal length, i.e., a chain of neighbouring explanations that all occur in an RT-window of finite, specified length ΔRT.

Considering the annotations individually as in the last section, we often find many valid explanations for any given peak. However, only one of these annotations will typically be the "true" solution, while the others occur by chance. The idea behind our refined procedure is then that false positive explanations will have a significantly smaller probability of supporting long charge ladders than the true positives.

The assignment of peak annotations into charge ladders can be achieved efficiently using a sweep algorithm [11]: for every protein i in the database, we iterate over all TMT-assignments $\beta_{i,j}$ that lead to a valid explanation. For each of those $(i, \beta_{i,j})$-pairs, we then determine a set of potential charge ladders, i.e., maximal sets of neighbouring assignments $\{(i, \alpha_{i,j} + k, \beta_{i,j}) | k = 0, \ldots, K\}$, disregarding the difference of their retention times (this can be done efficiently by sorting the set with respect to $\alpha_{i,j}$). Please note that, usually, many of the $(i, \alpha_{i,j} + k, \beta_{i,j})$ values will occur at different retention times.

Table 3 Results on the 10-protein mix from [6] when including the charge ladder filter with varying accuracy threshold ϵ, RT window ΔRT, and ladder length L.

ϵ	ΔRT	L	annotated peaks	valid annotations	found ladders	found proteins
10	10	2/3/4	52/10/0	54/10/0	24/3/0	6/3/0
10	20	2/3/4	71/15/0	77/15/0	30/4/0	7/3/0
10	50	2/3/4	83/18/0	91/18/0	31/5/0	7/3/0
20	10	2/3/4	253/83/9	269/87/9	117/28/2	7/5/1
20	20	2/3/4	343/149/44	371/161/56	151/52/10	7/6/2
20	50	2/3/4	408/186/84	444/206/100	151/50/17	7/7/3
100	10	2/3/4	3788/2823/2095	5657/4207/3128	1715/1034/640	8/8/7
100	20	2/3/4	4278/3674/3121	6405/5511/4701	1396/1033/774	8/8/7
100	50	2/3/4	4426/3955/3459	6621/5914/5181	792/571/434	8/8/7

In the next step, we consider each potential ladder individually to extract valid ladders within the given RT interval. For each $\alpha_{i,j}$-value in the current potential ladder, we store the RT values of all peaks that were explained by this annotation in a sorted list. Our sweep line starts at the annotation with lowest[4] RT value, regardless of its charge, and will progress in order of increasing RT value. In each step, we then try to extend valid ladders to the left and right, starting from the annotation currently touched by the sweep line, which will form the lower boundary of the RT interval ΔRT. The ladder can be iteratively extended if a neighbouring annotation within this RT interval exists. Since annotations are sorted with respect to RT, and since the sweep line always rests at the annotation with currently lowest RT value, it suffices to check the lowest-RT remaining (i.e., above the sweep line) annotation for each charge state. Thus, in each step, only one comparison per charge state is necessary. If a consecutive list of a user-defined minimal length L has been detected, all annotations in that list are marked as part of a charge ladder. Finally, the current annotation is removed from the list and the sweep line progresses to the next annotation. A snapshot at one step of the algorithm is depicted in Fig. 3.

For pre-sorted input, this algorithm obviously requires $\mathcal{O}(n \cdot L)$ operations, where n is the total number of annotations, since every annotation can be part of only one potential ladder and is touched at most once by the sweep line, and since every check requires up to L comparisons. Combined with the sorting steps, we arrive at a total runtime of $\mathcal{O}(n \cdot L + n \log(n))$.

3 Results of the refined procedure

We implemented the refined algorithm as a TOPP [9] tool, which is fully integrated into the OpenMS framework. The results of the refined procedure on the ten-protein mix are shown in Tab. 3. These demonstrate that charge-ladders indeed provide a strong filter: with increasing ladder length, the amount of remaining annotations drops particularly strongly for presumed low mass accuracy, and the amount of explanations per annotated peak becomes significantly smaller.

[4] In the degenerate case, where the minimum is not unique, annotations are visited from lowest to highest charge state.

In general, however, the problem setting is still highly ambiguous, despite the charge-ladder constraint. To further demonstrate this fact, we again applied our procedure on the data set of [6], but now with a much larger reference database to query: all proteins contained in UniProtKB/Swiss-Prot [2] that fall into a similar mass range as the ones known to be contained in the sample. The resulting data set consists of $3,990,159$ proteins, including the 10 true positives. Treatment of variable modifications doubles the database size to $7,980,318$ proteins. The computational efficiency of our method is demonstrated by the fact that the query of $16,042$ peaks against this huge database terminated in 488 minutes on a single core of a standard desktop PC (please note that the method can be trivially parallelized with nearly linear speed-up by splitting query database). However, the specificity on this data set is very low. Of the $7,980,318$ proteins in the database, $6,566,123$ have not only been annotated successfully, but also as part of stable charge ladders of length at least 3 in an RT-window of 20 seconds and with a maximal mass deviation of 20 ppm. If we choose more restrictive parameter values, some of the false-positive identifications indeed vanish (with $\epsilon = 10$ ppm the number drops to $1,112,502$), but so do the true positive ones.

4 Conclusion and Outlook

Top-down proteomics is a promising alternative to the popular shotgun approaches that are commonly applied. Its deficiencies in sensitivity are often made up for by avoiding coverage problems as they are common in bottom-up settings. Unfortunately, many of the established solutions for identification and quantification in the bottom-up domain cannot be simply transferred to the top-down case. This work was concerned with one such solution – the use of TMT-labelling for quantification purposes. In [6], the analytical background to apply TMT-markers to intact proteins was established, but resulted in a challenging annotation problem. In the pilot study, annotation was performed in a time-consuming manual fashion which can neither be performed on a high-throughput basis, nor generalized to the case of large reference databases.

Here, we have shown how the same annotation problem can be solved efficiently on a computer. Application of our method on the original data set has shown that the manual annotation was valid, but was only one of a variety of equally probable explanations. Only prior knowledge about the outcome allowed the annotator to select the "true" solution intuitively, which he validated by using further experimental constraints. We then proceeded to derive a refined algorithm for improved specificity without the need for further experimental effort. Through the use of so-called charge-ladders, we can exclude at least some of the noise present in the annotations, i.e., explanations that were only valid by chance without further supporting information in the spectrum. The resulting annotation problem seems to become significantly more complex, but can be solved efficiently using the sweep-paradigm.

Application to the data set has shown that the method is indeed efficient enough to be applied to real-world data sets. But whether it is specific enough to be applied routinely is still an open question. If at least partial sequence information is known from MS/MS experiments, the use of charge-ladders suppresses most false positive TMT/charge-variants. If ambiguity persists, all valid annotations will be returned to the user, who will then have to decide whether one of them can be trusted. Without such constraints, the amount of false-positives is clearly too large to be used routinely, even for moderate sample complexity. To go beyond the proof-of-concept stage in the general case will thus require at least an improved mass accuracy, but possibly also the use of other kinds of experimental constraints. This will be the focus of our future work.

Funding AH acknowledges financial support from the Intel Visual Computing Institute (IVCI) of Saarland University and the 'Schwerpunkt Rechnergestützte Forschungsmethoden in den Naturwissenschaften' of Johannes-Gutenberg University Mainz, AH and HPL financial support from DFG (BIZ4:1-4). AT received funding from the DFG Cluster of Excellence 'Inflammation at Interfaces'.

References

1 Marcus Bantscheff, Simone Lemeer, Mikhail M. Savitski, and Bernhard Kuster. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Analytical and Bioanalytical Chemistry*, 404(4):939–965, 2012.

2 The UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Research*, 40(D1):D71–D75, 2012.

3 Jimmy K. Eng, Ashley L. McCormack, and John R. Yates. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, 5(11):976–989, 1994.

4 Ross et. al. Multiplexed Protein Quantitation in Saccharomyces cerevisiae Using Amine-reactive Isobaric Tagging Reagents. *Molecular & Cellular Proteomics*, 3(12):1154–1169, 2004.

5 Simon J. Hubbard and Andrew R. Jones. *Proteome Bioinformatics*. Methods in Molecular Biology. Humana Press, 2010.

6 Chien-Wen Hung and Andreas Tholey. Tandem Mass Tag Protein Labeling for Top-Down Identification and Quantification. *Analytical Chemistry*, 84(1):161–170, 2012.

7 Rene Hussong and Andreas Hildebrandt. *Signal Processing in Proteomics*, chapter 11, pages 145–161. Methods of Molecular Biology. Humana Press, vol. 604 edition, 2009.

8 Rene Hussong, Andreas Tholey, and Andreas Hildebrandt. Efficient Analysis of Mass Spectrometry Data Using the Isotope Wavelet. In *COMPLIFE 2007: The Third International Symposium on Computational Life Science*, pages 139–49. American Institute of Physics (AIP) Proceedings Volume 940, 2007.

9 Oliver Kohlbacher, Knut Reinert, Clemens Gröpl, Eva Lange, Nico Pfeifer, Ole Schulz-Trieglaff, and Marc Sturm. TOPP - the OpenMS proteomics pipeline. *Bioinformatics*, 23(2):e191–e197, 2007.

10 David N. Perkins, Darryl J. C. Pappin, David M. Creasy, and John S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18):3551–3567, 1999.

11 Michael Ian Shamos and Dan Hoey. Geometric intersection problems. In *Foundations of Computer Science, 1976., 17th Annual Symposium on*, pages 208–215, 1976.

12 Martin Slawski, Rene Hussong, Andreas Tholey, Thomas Jakoby, Barbara Gregorius, Andreas Hildebrandt, and Matthias Hein. Isotope pattern deconvolution for peptide mass spectrometry by non-negative least squares/least absolute deviation template matching. *BMC Bioinformatics*, 13(1):291, 2012.

13 Andrew Thompson, Jürgen Schäfer, Karsten Kuhn, Stefan Kienle, Josef Schwarz, Günter Schmidt, Thomas Neumann, and Christian Hamon. Tandem Mass Tags: A Novel Quantification Strategy for Comparative Analysis of Complex Protein Mixtures by MS/MS. *Analytical Chemistry*, 75(8):1895–1904, 2003.