Published in final edited form as:

J Proteome Res. 2013 December 6; 12(12): . doi:10.1021/pr400705q.

Mass fingerprinting of complex mixtures: protein inference from high-resolution peptide masses and predicted retention times

Luminita Moruz[†], Michael R. Hoopmann[‡], Magnus Rosenlund[¶], Viktor Granholm[†], Robert L. Moritz[‡], and Lukas Käll^{¶,§,*}

Science for Life Laboratory, Department of Biochemistry and Biophysics, Stockholm University, 17165 Solna, Sweden, Institute for Systems Biology, 401 Terry Ave N, Seattle, Washington 98109, USA, Royal Institute of Technology - KTH, School of Biotechnology, Science for Life Laboratory, 17165 Solna, Sweden, and Royal Institute of Technology - KTH, Swedish e-Science Research Centre, 17165 Solna, Sweden

[†]Science for Life Laboratory, Department of Biochemistry and Biophysics, Stockholm University, 17165 Solna, Sweden

[‡]Institute for Systems Biology, 401 Terry Ave N, Seattle, Washington 98109, USA

[¶]Royal Institute of Technology - KTH, School of Biotechnology, Science for Life Laboratory, 17165 Solna, Sweden

[§]Royal Institute of Technology - KTH, Swedish e-Science Research Centre, 17165 Solna, Sweden

Abstract

In typical shotgun experiments, the mass spectrometer records the masses of a large set of ionized analytes, but fragments only a fraction of them. In the subsequent analyses, only the fragmented ions are used to compile a set of peptide identifications, while the unfragmented ones are disregarded. In this work we show how the unfragmented ions, here denoted MS1-features, can be used to increase the confidence of the proteins identified in shotgun experiments. Specifically, we propose the usage of *in silico* tags, where the observed MS1-features are matched against *de novo* predicted masses and retention times for all the peptides derived from a sequence database. We present a statistical model to assign protein-level probabilities based on the MS1-features, and combine this data with the fragmentation spectra. Our approach was evaluated for two triplicate datasets from yeast and human, respectively, leading to up to 7% more protein identifications at a fixed protein-level false discovery rate of 1%. The additional protein identifications were validated both in the context of the mass spectrometry data, and by examining their estimated transcript levels generated using RNA-Seq. The proposed method is reproducible, straightforward to apply, and can even be used to re-analyze and increase the yield of existing datasets.

Principle contribution—A statistical framework that uses the unfragmented MS1-features to increase the confidence of the proteins identified in shotgun experiments.

Introduction

Since its introduction in the late 1980's, peptide sequencing by mass spectrometry¹ has evolved to shotgun proteomics by the late 1990's,^{2,3} and has completely revolutionized the

* To whom correspondence should be addressed lukas.kall@scilifelab.se Phone: +46 (0)73707 8690. Fax: +46 (0)8524 81425.

Availability

The source code of the protein inference method we describe here is available under an MIT license at <http://code.google.com/p/insilico-amt/>.

way we conduct proteomics. The technique comprises the proteolytical digestion of the proteins in a complex biological mixture, the separation of the resulting peptides on a chromatographic column, and registering their mass-to-charge ratios and fragmentation spectra using a mass spectrometer. The current method to process this shotgun data is to first match the obtained fragmentation spectra against the theoretical spectra of all the peptides in a protein database, and subsequently infer proteins from the identified peptides. Normally, a mass spectrometer is operated in a way that it records the mass-to-charge ratios of all the analytes that were ionized sufficiently well, the so-called *MS1-features*, although it is capable of fragmenting only a subset of these analytes.

Currently, one of the main factors limiting the number of proteins that can be inferred from a mass spectrometry-based proteomics assay is the instrument's ability to fragment peptides.⁴ A theoretical digest of the human ENSEMBL v66 database comprises more than 6×10^5 unique tryptic peptides. To acquire one fragment spectrum for each of these peptides in a two hour experiment, one would have to detect 5000 peptides a minute, a figure that is far beyond the capabilities of the current instrumentation, which typically fragments just more than 400 analytes a minute.⁵ Post-translational modifications, inefficiencies of the enzymatic digestion, fragmentation in the ion-source, and possible contaminants further increase the complexity of the sample. Furthermore, the fragmentation events that are triggered by the on-board software of the mass spectrometer are selected based on ion abundance, leading to redundant sampling of the abundant peptides. As a consequence, a major direction to improve the yield of shotgun experiments is to collect more fragmentation spectra. This can be achieved either by augmenting the speed of the fragmentation mechanisms of the mass spectrometers,⁶ using improved setups in the chromatographic separation,^{7,8} or employing additional prefractionation techniques.⁹

Alternatively, in the absence of high confidence data supplied by the fragmentation spectra, one can use the observed retention times (RT) and mass determinations of the unfragmented analytes as additional input to the protein identification. This information is, for each individual peptide, less reliable than the evidence provided by a full fragmentation spectrum. This observation is, however, analogous to the individual ions in the fragmentation spectra themselves (Figure 1). Each separate ion in a fragmentation spectrum might not be unique enough to identify the right peptide from a database. Nevertheless, given an ensemble of ions of a fragmentation spectrum, we can often accurately select the correct peptide sequence. Likewise, individual matched MS1-features might not contain enough information to uniquely identify a protein, but the ensemble of such features can provide sufficient evidence to infer a protein.

MS1-features have traditionally been used as evidence for a particular protein in single protein experiments using mass fingerprinting. We can confirm the identity of a purified protein by investigating the correspondence between the expected peptides of a trypsinized protein and the observed masses.^{10–14} This technique, however, is not suitable for high-throughput studies, as it requires the proteins to be extracted, trypsinized, and analyzed one-by-one. In complex mixtures accurate mass and time tags (AMTs) have been used as means of peptide identification.^{15,16} The combination of the mass and retention time of a peptide identified by fragmentation in a prior experiment is recorded, and presence or absence of a similar observation in subsequent experiments is seen as evidence for the presence or absence of that particular peptide. However, the AMT method has the drawback that the peptide tags need to be accurately identified in a prior experiment in order to record their retention time.

An alternative strategy is to define *in silico* AMTs by predicting the retention time of the theoretical peptides *de novo*. Such *in silico* AMTs are then matched to unfragmented MS1-

features forming *peptide-feature matches* (PFMs). Previously, PFMs have been used to infer protein sequences in lower level organisms.^{17,18}

Here, we show that the PFMs can be used as additional input to the task of identifying proteins, and propose a statistical framework to compute protein probabilities using both fragmented and unfragmented ions. We applied our method for two triplicate datasets of complex peptide mixtures, and showed that our approach provides additional confidence to the proteins identified using the fragmentation spectra, while leading at the same time to up to 7% more protein identifications at a fixed protein-level false discovery rate of 1%. We validated the additional proteins both in the context of our mass spectrometry data, and by inspecting their corresponding transcript levels obtained from independent experiments. In terms of reproducibility, our approach is comparable to the typical workflow based solely on the fragmentation spectra. We conclude by discussing the potential of using PFMs to distinguish protein homologs in shotgun studies, and for other applications in mass spectrometry-based experiments.

Experimental

Sample preparation

Yeast strain BY4742 (haploid mating type alpha) with NUP192 protein A tag was obtained as a gift from the Aitchison Lab (Institute for Systems Biology). The cultures were grown to mid-log phase and harvested by centrifugation. The cells were lysed by flash freezing in liquid nitrogen prior to disruption using a Retsch ball mill grinder and resuspended in buffer containing 8M urea and 100 mM ammonium bicarbonate. Proteins were denatured with 5 mM TCEP, and free sulfhydryl bonds were alkylated with 5 mM iodoacetamide. The proteins were digested to peptides by incubation with trypsin for 16 hours at room temperature. The pH was adjusted to approximately 2 by addition of TFA.

Human Du145 prostate cancer cells were washed in cold PBS, and lysed in lysis buffer (8M urea, 0.1% rapigest (Waters, USA), 100mM ammonium bicarbonate). Once lysed, the sample was diluted 8-fold with 100mM ammonium bicarbonate and protein concentration was measured by BCA assay. The proteins were denatured with 5 mM TCEP and free sulfhydryl bonds were alkylated with 10 mM iodoacetamide. The proteins were digested with trypsin overnight. HCl was added to a final concentration of 50 mM and TFA was added to a final concentration of 1%. Peptides were desalted using C18 spin columns.

LC-MS/MS analysis

LC-MS/MS analysis was performed using a IntegraFrit (New Objective, USA) capillary (75 μ m ID) packed with 20 cm of ReproSil Pur C18-AQ 3 μ m beads (Dr. Maisch GmbH, Germany), and joined by union to a PicoTip (New Objective, USA) pulled silica tip (20 μ m ID). Prior to loading the column, the sample was loaded onto a fritted capillary trap (75 μ m ID) packed with 2 cm of the same material. For each sample injection, 1 μ g total protein was loaded onto the trap using an Agilent 1100 binary pump. Each sample was separated using a binary mobile phase gradient to elute the peptides. Mobile phase A consisted of 0.1% formic acid in water, and mobile phase B consisted of 0.1% formic acid in acetonitrile. The gradient program consisted of three steps at a flow rate of 0.3 μ L/min using an Agilent 1100 nanopump: (1) a linear gradient from 5% to 40% mobile phase B over two hours, (2) a 10 minute column wash at 80% mobile phase B, and (3) column re-equilibration for 30 minutes at 5% mobile phase B.

Mass spectra were acquired on a LTQ Velos Orbitrap (Thermo Fisher Scientific) mass spectrometer operated on an 11-scan cycle consisting of a single high-resolution precursor scan event at 60,000 resolution (at 400 m/z) followed by 10 data-dependent MS/MS scan

events using collision induced dissociation (CID). The data-dependent settings were a repeat duration of 30 seconds, a repeat count of 2, and an exclusion duration of 3 minutes. Charge state rejection was enabled to fragment only 2+ and 3+ ions. Additional parameters included the mass range (MS1) set to 400-1400 m/z, AGC on, 1e6 ions, lock mass off, normalized fragmentation energy (MS2) set to 35.0, and the isolation width for MS2 set to 2.0.

All the raw files are available online via <http://www.nada.kth.se/~lumi/datasets/iamt/iamt.html>

Data processing

The tool MakeMS2 v2.28 with default parameters was used to convert the raw data to the .ms1 and .ms2 file formats. The .ms1 data was subsequently deconvoluted and the monoisotopic mass and charge state of the analytes were determined using Hardklör v2.01.¹⁹ Next, we used Krönik v2.02²⁰ to determine those features that were observed in at least five consecutive scans, with a gap tolerance of one scan. The parameter values used to run Hardklör and Krönik are listed in the Supplementary Data, as Supplementary Information 1. MakeMS2 can be downloaded at <http://proteome.gs.washington.edu/software/makems2/>.

The fragmentation spectra were searched with Crux v1.35,²¹ using the sequest-search command. The only modification searched for was the carbamidomethylation of cysteine (static modification of 57.021464 Da to all cysteines). We used a precursor mass window of 10ppm, the enzyme set to trypsin, and no missed cleavages were allowed. The parameter values used to run Crux are listed in the Supplementary Information 1. The three yeast datasets were searched against the ENSEMBL v64 protein database (6,696 entries), while for the human datasets we used Swissprot 2011_09 (20,185 entries). All the datasets were also searched against a decoy database obtained by reversing the protein sequences from the target database. We used Bullseye v1.3,²² with default parameters, to assure that the retention time of each peptide was assigned to the apex intensity of its corresponding feature.

The resulting datasets were post-processed using Percolator v2.01,²³ which provided posterior error probabilities and false discovery estimates at the peptide level. Protein posterior probabilities were calculated using Fido,²⁴ and *q* values were calculated by averaging posterior error probabilities. Table 1, first three columns, gives the number of peptides and proteins identified at a false discovery rate of 1%.

The full lists of peptide and protein identifications are accessible online at <http://www.nada.kth.se/~lumi/datasets/iamt/iamt.html>.

Mass calibration and mass error estimation

The peptides longer than ten amino acids confidently identified from the fragmentation spectra (*q* < 0.01) were used to improve the mass accuracy of the MS1-features. A mass recalibration algorithm was written that accepts MS1 spectra and a list of identified peptides. Between 88% and 90% of the identified peptides were matched by mass (< 10ppm) and retention time (± 15 s) to peptide features in each MS1 spectra, and a mass difference was computed from the observed and expected peptide masses. Although peptides are identified from a single scan event, their precursor ions are observed over many consecutive MS1 scan events, and thus it is possible to map single peptide identifications to multiple MS1 spectra. Therefore, many of the MS1 spectra contained several peptide matches that showed systematic trends in mass accuracy across the m/z scan range. For spectra that had two or more peptide matches, a linear line of best fit was computed for the mass differences versus

m/z values. The best fit function was then applied as a correction to the m/z values of all data points in the spectrum. The mass recalibration software can be downloaded at <http://code.google.com/p/rawaccuracy/>.

After applying the calibration algorithm, we ran Hardklör and Krönik once more to obtain a list of calibrated MS1-features. The differences between the theoretical masses of the peptides longer than ten amino acids identified from the fragmentation spectra, and the masses of the corresponding calibrated MS1-features, were used as an estimate for the distribution of the mass errors in our data. Supplementary Figure 1 illustrates the distribution of all these differences before and after applying the recalibration algorithm.

Furthermore, to increase the confidence in the MS1-features, we retained only the 50% of the features with the highest intensity. Table 1, fourth column, displays the final number of MS1-features for each dataset.

Retention time prediction

We estimated the retention time of the peptides from an *in silico* digest of the proteome using ELUDE.²⁵ Since ELUDE is a machine learning-based method, it first needs to train a retention model using a set of accurate peptides with known retention times. In this context, we used two thirds of the peptides longer than ten amino acids identified from the fragmentation spectra ($q < 0.01$) to train a retention model, and the remaining one third to evaluate the performance of the model.

Supplementary Figure 2 illustrates the predicted retention time as a function of the observed retention time for the peptides used to evaluate the models. Since in the yeast runs the error of the predictions for peptides with retention time above 100 minutes was significantly higher (Supplementary Figure 2, A-C), we chose to retain for subsequent analyses only the MS1-features eluting in the first 80% of the run.

Peptide-feature mapping

We performed *in silico* digestions of the yeast and human proteomes, and kept all peptides with masses ranging from 600 to 8000 Da that were between 11 and 50 amino acids long. From the resulting sets, we removed all peptides that were matched to a fragmentation spectra with a posterior error probability less than 1.0. This ensured that the information obtained from the MS1-features was fully independent from the fragmentation spectra. In addition, to avoid ambiguities, we have also excluded those peptides that mapped to more than one protein.

Here it is worth emphasizing that we did not remove from the analysis the MS1-features corresponding to the peptides matched to fragmentation spectra. This is because we still expected a large proportion of the peptides identified at posterior error probability below 1.0 to be incorrect matches. By keeping all the MS1-features in the analysis, these peaks got the possibility to be matched to the correct peptide sequences.

Next, for each of the *in silico* peptides, we computed the theoretical monoisotopic mass, and estimated its retention time using a retention model trained as previously described. The joint distributions mass/retention time for the theoretical peptides, as well as for the MS1-features are displayed in Supplementary Figure 3.

In the following step, we matched the list of observed MS1-features to the *in silico* peptides based on mass and retention time, allowing for a fixed tolerance window as explained below. A pair consisting of a theoretical peptide and an observed MS1-feature was

considered a peptide-feature match (PFM) if it fulfilled $\varepsilon < r$, with r being a predefined threshold, and the combined normalized error ε defined through:

$$\varepsilon = \sqrt{\frac{(\delta_m - \bar{\delta}_m)^2}{\sigma_m^2} + \frac{(\delta_t - \bar{\delta}_t)^2}{\sigma_t^2}} \quad (1)$$

where δ_m is the mass difference between the peptide and the MS1-feature, and $\bar{\delta}_m$ and σ_m are the average and standard deviation of the mass errors for the peptides identified from the fragmentation spectra, respectively. Similarly, δ_t is the difference in retention time between the peptide and the feature, and $\bar{\delta}_t$ and σ_t are the average and standard deviation of the retention time prediction errors for the confidently identified peptides, respectively. The threshold r controls the maximum permitted difference in mass and retention time between the MS1-feature and the peptide of a PFM.

To set the threshold r , we first calculated the combined normalized error ε for each of the peptides identified from the fragmentation spectra that could be uniquely mapped to the corresponding MS1-feature. We then set r to be the smallest ε value that encompassed 80% of these peptides. More intuitively, if we would plot the joint distribution of the standardized mass and retention time errors for the identified peptides, our procedure uses the 80% of the data points closest to the center to define a circle of radius r (Supplementary Figure 4). With such a representation, the PFMs will be those pairs (theoretical peptide, MS1-feature), whose mass and retention time differences fall within this circle.

Since a peptide can match more than one feature, and conversely, a feature may match more than one peptide, a PFM does not necessarily correspond to a unique peptide or feature.

Table 1 displays the values of $\bar{\delta}_m$, σ_m , $\bar{\delta}_t$, σ_t and r , as well as the total number of PFMs, for each of the investigated datasets.

Protein inference from PFMs

Frequently peptides do not match uniquely to a feature (and vice-versa), and therefore we may assume that a non-negligible proportion of the detected PFMs are incorrect. In this context, it is essential to build a statistical method that can allow us to express that a protein is present given the observed PFMs.

The first step in this procedure was to define a null model, *i.e.* a model describing the distribution of the PFMs for incorrect proteins. For this we ordered all the peptides from the *in silico* digest according to their monoisotopic mass, divided them in bins of about 120 peptides each, and then shuffled the predicted retention times within each bin. We subsequently matched the peptides to the observed MS1-features using their mass and the new retention time following the procedure previously described. We assumed that all the resulted PFMs were random matches. Further, for each length l , we could estimate the probability $q(l)$ that a peptide of length l would be randomly matched to MS1-features by the fraction of all peptides of length l matched to MS1-features in the null model (Supplementary Tables 1 and 2). We then used these probabilities to assign protein p values using a dynamic programming procedure as described below.

Let R be a protein having a set of n constituent peptides. Each peptide i has the probability $q_i = q(l_i)$ to be incorrectly matched to MS1-features, where l_i is the length of the peptide i . The results of matching the peptides of the protein R to MS1-features can be expressed as an array $D = (d_1, \dots, d_n)$, where $d_i = 1$ if the constituent peptide has a matching MS1-feature, and

$d_i = 0$ otherwise. We can then calculate the probability that the matches D occurred by chance as,

$$\Pr(D|R=0) = \prod_{i:d_i=1} q_i \prod_{i:d_i=0} (1 - q_i) = \prod_{i=1,\dots,n} q_i \prod_{i:d_i=0} \frac{1 - q_i}{q_i}. \quad (2)$$

Given that we observe D for protein R , we are interested in calculating the chance of observing a PFM configuration D' at least as extreme as D . This translates into calculating the summed probability, $\sum_{D'} \Pr(D'|R=0)$ of permutations D' such that $\Pr(D'|R=0) \geq \Pr(D|R=0)$. By studying the logarithm of Equation 2, we see that this is equivalent to finding the summed probability of the permutations that fulfill the criterion,

$$\sum_{i:d_i=0} x_i \leq y,$$

where $x_i = \log\left(\frac{1 - q_i}{q_i}\right) = -\text{logit}(q_i)$, and $y = \log(\Pr(D|R=0)) - \sum_i \log(q_i)$.

An approximation to the problem can be obtained by discretization, scaling and rounding each x_i to an integer value such that $X_i = \text{round}\left(\frac{x_i}{k}\right)$, where k is a user-defined, sufficiently small, scaling factor. In our calculations, we used $k = \max_i(x_i)/1000$. Note that since all $q_i \leq 0.5$, the variables X_i are always positive integers. With the new representation of the problem, we need to search for the permutations that fulfill:

$$\sum_{i:d_i=0} X_i \leq Y,$$

where $Y = \text{round}\left(\frac{y}{k}\right)$. Let us also introduce a variable $Q = \sum_{i=0}^n \log(q_i)$. Studying Equation 2, we can use this new set of variables we can express an approximation of the logarithm of the probability of a given permutation D as,

$$\log(\Pr(D|R=0)) = \log\left(\prod_{i=1,\dots,n} q_i \prod_{i:d_i=0} \frac{1 - q_i}{q_i}\right) = Q + \sum_{i:d_i=0} x_i \approx k \sum_{i:d_i=0} X_i + Q. \quad (3)$$

Let $f(s, j)$ be a function that expresses the number of permutations $d_1, \dots, d_j$, such that $\sum_{i:d_i=0} X_i = s$. Then for $j = 1, \dots, n$, we have,

$$f(s, j) = \begin{cases} f(s, j-1) + f(s - X_j, j-1) & \text{if } s \geq X_j \text{ and } j > 0 \\ f(s, j-1) & \text{if } s < X_j \text{ and } j > 0 \\ 1 & \text{if } s=0, j=0 \\ 0 & \text{if } s > 0, j=0. \end{cases}$$

We can efficiently calculate $f(s, n)$ by first allocating a vector F of length $S = 1 + \sum_{i=0}^n X_i$ with all the elements set to zero, except for the first element which is set to one. From the end of the vector we then add the content of the element X_j indexes lower down to the

content of the current element, and walk down until the beginning of the vector. We repeat the procedure for $j = 1, \dots, n$.

Finally, to derive an approximation of the p value for our protein R given the observations D , we calculate the probability $p_R(D)$ of all permutations with scores at least as extreme as D using the approximation of Equation 3,

$$p_R(D) = \sum_{D': \Pr(D'|R=0) \leq \Pr(D|R=0)} \Pr(D'|R=0) \approx \sum_{s \leq Y} f(s, n) \cdot e^{ks+Q}. \quad (4)$$

This procedure was used to calculate a set of accurate p values for all the proteins in our datasets. To correct for multiple testing, we subsequently calculated the corresponding q values and posterior error probabilities using qvalue.²⁶

Statistical calibration of the protein-level PFM-derived p values

For continuous distributions one can easily confirm the calibration of the p values by verifying whether the p values generated under the null hypothesis are uniformly distributed in the interval $[0, 1]$. However, the p values generated from discrete distributions will be a discrete function due to the limited number of outcomes. To still be able to assess the calibration of our method, we calculated so-called “randomized” p values. This was done for the purpose of the calibration, by replacing Equation 4 by,

$$\widetilde{p}_R(D) = a \cdot f(s, Y) \cdot e^{kY+Q} + \sum_{s < Y} f(s, n) \cdot e^{ks+Q}, \quad (5)$$

where a is a randomly generated number in the interval $[0, 1]$. Intuitively, this method is equivalent to dealing with ties when computing empirical p values by randomly spreading the probability of the permutations at the threshold between the rejection region and the non-rejection region.

The advantage of using this approach is that the p values for the truly null hypotheses are smoothed out, becoming uniformly distributed in the interval $[0, 1]$. To test that this was indeed the case for our data, we built for each protein in the target database an incorrect counterpart by randomly choosing peptides present in other proteins (Supplementary Algorithm 1). Note that each such random-peptide containing protein had the same number of peptides as its non-decoy analogue, and that our algorithm preserved the number of peptides shared between different proteins. Next, we computed for each such protein the number of peptides that matched to a feature according to the null model, and assigned p values as described above. Figure 2A, illustrates the distribution of the resulting p values for the first yeast dataset, while in Supplementary Figure 5 we depict the same p values against hypothetical, perfectly uniform p values. As expected, since the random-peptide containing proteins correspond to truly null hypotheses, we obtain a uniform distribution for the p values. For a comparison, Supplementary Figure 6 illustrates similar plots when calculating exact p values. These results were replicated for all the datasets under study.

The randomized p values were used as a way to assess the calibration of the statistical results yielded by our method. When calculating PFM-based protein probabilities, we used the more conservative p -values obtained using Equation 4.

Combined protein level probabilities

We used the approach described above to assign to each protein a PFM-based posterior error probability. Since the peptides matching a fragmentation spectrum with PEP < 1.0 were

discarded when deriving the PFMs, the PSM- and PFM-based probabilities were completely independent from each other. As a consequence, we chose a simple scheme to compute combined probabilities, by multiplying for each protein the PFM-based PEP with the PEP computed using the fragmentation spectra. Next, we computed corresponding q values by averaging the PEPs. The full list of protein identifications, together with their PSM+PFM-derived statistics are available online at <http://www.nada.kth.se/~lumi/datasets/iamt/iamt.html>.

Validation of protein identifications using transcript data

For the three human datasets, we have also investigated the abundances at the transcript level for the proteins identified using our *in silico* AMT method. For this, we downloaded the dataset corresponding to a high-throughput sequencing of polyA+ RNA (RNA-Seq) experiment for the Human DU145 cells (data accessible at the NCBI GEO database,²⁷ accession GSE25183).²⁸ We used Cufflinks v2.02²⁹ with hg18³⁰ as reference annotation to derive the estimated expression level for each gene. The expression levels were expressed in FPKM (Fragments Per Kilobase of transcript per Million mapped reads), and we included only those genes for which the deconvolution was successful (FPKM_status OK). Further, we used the hgncXref table available from the UCSC website³¹ to translate the gene names to corresponding UniProt identifiers. Finally, we matched the resulting list of expression levels to the protein identifications found with the *in silico* AMT method. Whenever a protein mapped to more than one gene, we considered the average FPKM value of all its corresponding genes.

In addition, Dr. Maud Starmans kindly provided us transcriptome profiling data generated using gene-expression microarrays for the same Human DU145 cells. Details of the experimental setup are described in.³² We used the batch extract interface for SOURCE at <http://smd.stanford.edu/cgi-bin/source/sourceBatchSearch> to find the corresponding protein identifier for each gene, and as previously, we estimated the expression level of each protein as the average of the expression values of its corresponding genes.

Results

An *in silico* AMT workflow

The method we propose in this work - denoted *in silico* accurate mass and time tag (*in silico* AMT) - includes two parallel workflows: one based on the fragmentation spectra, and one centered around the MS1-features (Figure 3). The details of each of these workflows are given in the Experimental procedures section. Briefly, for the fragmentation spectra we employ Crux and Percolator to identify a list of accurate peptide identifications, and Fido to translate these to protein probabilities. For the workflow based on the MS1-features, Hardklör and Krönik are used to compile a list of reliable MS1-features. We calibrate the observed masses of the MS1-features and estimate their mass accuracy using an in-house algorithm trained with the peptides confidently identified from the fragmentation spectra. Further, we perform an *in silico* digest of a protein database, and retain only those peptides that are unique to one protein and that do not match to a fragmentation spectrum. We compute the theoretical masses of these peptides and employ our retention time predictor, ELUDE, to estimate their retention times. Observed MS1-features are then matched to the theoretical peptides by mass and retention time. Any match between a theoretical peptide and an MS1-feature is denoted a peptide-feature match. We build a null model by randomizing the predicted retention times of the theoretical peptides, and use this to compute PFM-based protein probabilities. In the final step of the workflow, we combine for each protein the probability derived from the fragmentation spectra with the one computed using the MS1-features.

***In silico* AMT detects a substantial fraction of the proteins identified using fragmentation spectra**

We first validated our *in silico* AMT approach by investigating whether it can detect the proteins that already had compelling evidence of being present from the fragmentation spectra. For this, we considered the full *in silico* digest of the yeast and human proteomes, including the peptides confidently matched to fragmentation spectra. We then used our *in silico* AMT method to match these peptides to the MS1-features, and to compute protein probabilities based on the resulting peptide-feature matches. Finally, we compared the confident protein identifications yielded by this approach, based solely on the MS1-features, with the proteins identified using the typical workflow based on the fragmentation spectra.

Our results showed that between 436-471 (90-92%) for yeast, and 786-797 (90-91%) for human, of the proteins found at 1% FDR using *in silico* AMT were also identified from the fragmentation spectra at the same FDR threshold. These results clearly indicate the usefulness of the MS1-features for protein identification, and show the efficacy of our method in translating this data into statistically meaningful information. In addition, an interesting observation was that using the MS1-features alone within the *in silico* AMT framework, we were able to detect about half of the proteins identified from the fragmentation spectra at the same FDR cutoff (Supplementary Table 3).

Protein inference from PFM-data

Since our goal was to use the MS1-features as additional evidence for the proteins identified from the fragmentation spectra, we removed from the *in silico* digest all the peptides that were matched to a fragmentation spectrum or mapped to more than one protein. We then applied once more our *in silico* AMT workflow as described in the Experimental Procedures section, and derived PFM-based protein statistics. Figure 2B illustrates the distribution of the p values computed for all the proteins present in the yeast database for yeast-01, and Supplementary Figure 7 illustrates the results obtained for the other datasets. As expected, the distribution displays an enrichment of low p values corresponding to proteins for which we have strong evidence of being present in the sample, and it evens out at larger p values, as these will mostly correspond to incorrect protein identifications. Subsequently, to correct for multiple testing, we used the software *qvalue*³³ to derive the corresponding PFM-derived q values and posterior error probabilities (PEPs) from the p values.

Notably, after removing the peptides matched to the fragmentation spectra from the analysis, the number of proteins confidently identified by PFMs decreased dramatically (Supplementary Table 4). Nevertheless, even if all peptides matched to MS2 spectra were removed from the analyses, we still found that between 62-83% of the proteins identified at PFM-derived $q < 0.01$ in the yeast datasets were also identified from the fragmentation spectra at the same q value threshold. Similar results were obtained for the human datasets, where 85-87% of the proteins identified at PFM-derived $q < 0.01$ were also confidently identified from the fragmentation spectra.

Improving protein level statistics using PFM-based information

Since the PFM-derived results were obtained independently from the fragmentation data-derived results, we modeled the errors from the two processes as being independent from each other. Hence, we could calculate the combined protein-level statistics for a protein by multiplying the PSM-derived PEP computed by Fido with the PFM-derived PEP calculated by our *in silico* AMT approach. By using such a procedure, we only employ the PFMs to supply additional confidence to the protein identifications generated from the fragmentation spectra, rather than diluting the confidence provided by the fragmentation data. More clearly, the PEP assigned to a protein will never increase after incorporating the PFMs.

However, the PEPs of proteins that have strong support from the PFM s may decrease after including this information, and thus such proteins may rank higher than other having only PSM-based evidence.

As expected, most of the proteins that obtained low PEPs after incorporating the PFM information were already identified from the fragmentation spectra, and therefore the PFM s provided additional confidence that these proteins were actually present in the sample. Furthermore, by combining the two probabilities, we increased the number of proteins identified at a fixed q value. Table 2 gives the number of additional proteins obtained at $q < 0.01$ for each of the six datasets investigated, while Supplementary Tables 5-10 give the full list of new protein identifications obtained after including the PFM s. For yeast, we obtained between 5.8% to 7.4% more proteins, while for the human samples the same numbers ranged between 5.4-5.7%. If we lowered the threshold to $q < 0.05$, we obtained up to 9 and 7% more proteins for the yeast and human datasets, respectively. Figure 4A illustrates the number of proteins at different q value thresholds obtained for the yeast-01 dataset when using the PSM s, and when combining the PSM s and PFM s. Similar representations are given for the other datasets in Supplementary Figure 8.

These increases are similar to the ones obtained from the fragmentation spectra when combining the three replicates of each dataset. As an example, when using percolator²³ and fido²⁴ to combine the three yeast replicates, we obtained 4453 peptides and 1051 proteins at 1% FDR. These translated to average increases of 19% and 7% at peptide and protein level, respectively. For the human datasets the three replicates together gave 7983 peptides and 1754 proteins, corresponding in average to 19% more peptides and 6% more proteins than identified in one single run.

In addition, we have investigated for how many of the additional proteins identified at $q < 0.01$ we were able to find peptides matched as best hits to a fragmentation spectrum. Our data showed that more than 98% of the additional proteins identified for yeast, and 97% for human, had at least one peptide matched to a fragmentation spectrum. This implies that for yeast 2% of the additional identifications found at 1% FDR had support solely from PFM s. In absolute numbers, this is equivalent to one protein. For the human datasets, 3% of the additional identifications, corresponding to about 3 proteins, were confidently identified from PFM s alone.

Reproducibility of the identifications

Since our data consisted of triplicate runs for both yeast and human, we were able to evaluate the reproducibility of our *in silico* AMT workflow by investigating the overlap between the proteins identified in the three datasets. For this we considered all the protein identifications with PSM+PFM-derived $q < 0.01$, and checked for each dataset how many of these identifications were found in the other two replicate runs (Figure 4B). Following this procedure, we found that between 94-95% of the proteins identified in one of the yeast datasets were identified in at least one of the other two runs at the same q value threshold. Comparable results were obtained for the human data, where the same numbers ranged between 91-94% (Supplementary Figure 9B).

As a comparison, we have also examined the reproducibility of the protein identifications when using the PSM s alone. When checking the overlap between the proteins identified at PSM-derived $q < 0.01$, we found that between 93-95% of the proteins identified in one of the yeast runs at PSM-derived $q < 0.01$ were also identified in at least one of the other two datasets at the same confidence threshold (Figure 4C). The results were replicated for the human data, where these numbers ranged between 91-95% (Supplementary Figure 9A).

In addition, we have also performed similar analyses at the peptide level. This time we considered all the peptides matched to fragmentation spectra or to MS1-features that belonged to proteins with PSM+PFM-derived q -values below 1%, and checked the overlap between these peptides across the three replicates (Supplementary Figures 10B, 10D). Our results showed that between 88-90% of the yeast peptides, and between 87-88% of the human peptides that were found in one dataset were replicated in at least one of the other datasets. These figures were similar to the ones obtained when using the PSMs alone, where the same numbers ranged between 89-90% for yeast, and 87-90% for human (Supplementary Figures 10A, 10C).

Thus, both at peptide and protein level, the reproducibility of the *in silico* AMT method was comparable to the one obtained when using solely the fragmentation spectra.

Validation of additional proteins

By the virtue of its design, our method aims at employing the PFMs to *enhance* the confidence of the protein identifications derived from the fragmentation spectra. As a consequence, if we consider a q value threshold of 0.01, virtually all the proteins that were identified from the fragmentation spectra at this threshold were also present in the final list obtained after incorporating the PFMs. Figure 5A illustrates a typical example of a human protein that had both PSM-derived and combined PSM+PFM-derived q values below 0.01. While the information supplemented by the PFMs is not essential for identifying such proteins, they do provide valuable information and allow us to compute more accurate statistics at the protein level.

More interesting are however the additional proteins identified using our *in silico* AMT method. We have divided these proteins in two groups:

- Proteins that were *nearly-identified* from the fragmentation spectra, *i.e.* proteins that had a PSM-derived q value between 0.01 and 0.05.
- Proteins that were not identified from the fragmentation spectra, *i.e.* proteins with a PSM-derived q value above 0.05.

From the new proteins identified at $q < 0.01$ in the yeast datasets, between 67-74% were nearly-identified from the fragmentation spectra, while the same numbers ranged between 46-63% for the triplicate human datasets. An example of such a protein for the human-01 data is displayed in Figure 5B. For such cases the PFM information is essential, as it allows us to select from the proteins with a score near the cutoff those identifications that are most probable to be correct. Furthermore, for the yeast data, between 42-53% of such proteins were identified from the fragmentation spectra in at least one of the other yeast datasets at PSM-derived q value below 0.01. The same numbers ranged between 43-60% for the human data.

Finally, the most challenging examples are those from the second category described above, for which we have rather weak support from the fragmentation spectra, but very strong evidence from the PFMs. Figure 5C illustrates one of the most extreme such cases for the human-01 dataset, where most of the evidence of a protein was provided by the PFMs. Even for these cases, we found that in the yeast data between 20-53% of these proteins were identified in at least one of the other datasets at PSM-derived $q < 0.01$. The same figures were lower for the human data, where between 27-41% of these proteins were confidently identified at PSM-derived $q < 0.01$ in another human dataset.

In addition, we have investigated the mass errors obtained by the PFMs. Supplementary Figure 11 displays the distributions of these errors for the PFMs mapping to the proteins

confidently identified from the fragmentation spectra, as well as for the PFMs mapping to the additional proteins identified with our *in silico* AMT method. As expected, the two distributions are similar to each other, and to the distribution of the mass errors observed for the confident peptide-spectrum matches.

Comparison with transcript data

To further validate our findings, we have also examined the mRNA abundances for the proteins identified using our *in silico* AMT approach. For this, we have processed a publicly available dataset generated using high-throughput sequencing as described in the Experimental procedures section, and obtained estimated transcript-level abundances for our protein identifications. Figure 6 displays the distribution of the abundance levels for the proteins identified at PSM-derived $q < 0.01$, and the estimated mRNA levels corresponding to the additional proteins obtained after including the PFMs. Supplementary Figure 13 gives similar representations when examining the transcript levels measured using microarray data.

As anticipated, most of the identified proteins correspond to highly abundant transcripts. This observation is valid for both the confident proteins identified from the fragmentation spectra, and the additional proteins found using *in silico* AMT. On average, the PFM-additional proteins seem to correspond to slightly lower transcript levels. This is nevertheless expected, as these proteins are inferred from less abundant ions than the PSM-confident proteins. This latter observation is even more apparent when inspecting the MS1 intensities for PSMs and PFMs (Supplementary Figure 14).

Discussion

Here we have shown that, by mapping the estimated masses and retention times of theoretical peptides to observed MS1-features, we can increase the confidence of identified proteins in a mass spectrometry-based proteomics experiment. Hence, we have demonstrated an additional route to increase the yield of an experiment by using information that, although readily available, is typically overlooked in the data analysis workflows.

Our *in silico* AMT method can be easily applied to any shotgun experiment, the only requirements being a high resolution precursor feature detection and the availability of retention time predictors for the LC system employed, which is the case for a majority of the mass spectrometry-based proteomics experiments. This implies that the technique can be used to re-mine existing datasets fulfilling these requirements, and thus obtain new and improved results from data that was previously discarded from the analysis. Consequently, we are validating our approach not only for future endeavors, but also for experiments conducted as far back as eight years ago. In addition, new mass spectrometry data intensive workflows based on Data-independent analysis (i.e., DIA or SWATH-MS³⁴) can take advantage of the *in silico* AMT approach through the additional analysis of multiplexed data collected in a high-reproducible and comprehensive data collection mode.

Currently, comparisons of different variants of the shotgun proteomics workflow are frequently expressed in terms of number of PSMs at a certain FDR threshold. Here we are, due to the PSM-free nature of parts of our data, unable to produce such comparisons. Nevertheless, since our method is focused on the protein identification problem, protein-level comparisons are more natural. Furthermore, the majority of the confidently identified PSMs of a shotgun proteomics experiment map to a relatively low number of proteins. To illustrate this, we displayed a cumulative comparison of the number of identified proteins and PSMs for the yeast dataset (Figure 7). This data shows that more than 50% of the PSMs,

stems from only 10% of the identified proteins. Hence, even low augmentations in number of identified proteins, translate to relative large increases in number of identified PSMs.

We would like to stress that the workflow described in this study should be regarded as a proof of principle rather than a definite solution, with particularly stringent choices at each step. One can imagine that the approach will have an even larger impact if we include all the MS1-features, not only the abundant ones, all the theoretical peptides, without excluding those mapping to a fragmentation spectrum. Regarding the latter observation, it is noteworthy that the scheme to combine the two probabilities used in this work is rather simple. One can imagine scenarios where more complex methods to combine the two types of information can be used. Additionally, with the availability of retention time predictors for post-translationally modified peptides,³⁵ the procedure can be extended to include such peptides as well. Also, allowing for missed cleavage site is one more future extension of the current workflow.

Our workflow, as described in the current study, makes use of accurate retention time predictions for the theoretical peptides in a proteome. Nevertheless, there are a number of chromatography systems for which no such predictors are currently available. In such cases, the method cannot be applied in the form described here. However, since the key point in achieving specificity is the use of high accuracy masses, one may still attempt to match the MS1-features to theoretical peptides based on mass alone. Such an approach would generate more non-specific matches, and thus is expected to lead to smaller improvements in terms of additional protein identifications. A similar effect is expected from drops in mass or retention time prediction accuracies. Our method models these two error types jointly, and thus it is not meaningful to try to calculate the precise influence of any of them separately. Furthermore, the extent of such effects are expected to vary with instrumentation and complexity of biological sample.

The method described in this study shows how, by incorporating the additional information provided by the unfragmented ions, we can increase the output of a shotgun proteomics experiment. Nevertheless, the MS1-features represent just one type of additional information that can be included in our analyses. Retention times, ion intensities and isotopic distributions are additional examples of data that could be used in a similar fashion. Furthermore, as much as 80% of the fragmentation spectra generated in a typical experiment may remain unexplained.³⁶ All these suggest that, besides the obvious focus on improving the instrumentation, an equally important concern should be directed towards designing efficient computational and statistical methods to integrate and use the wealth of information generated in each experiment.

Critical evaluation

Our proposed algorithm uses the collection of peptides matched to MS1-features to assign protein probabilities. The focus here is on the protein inference, and not peptide inference. Hence, each individual PFM may be unreliable, and should not be regarded as an identified peptide in itself. This is analogous to the fragmentation spectra (Figure 1), where we are interested in the overall confidence of a peptide identification, rather than the individual fragments of that peptide.

Furthermore, our workflow includes a number of parameters such as the percentage of high intensity MS1-features considered, the cutoff used to define a peptide-feature match, or the size of the bins when defining the null model. While our framework did not seem particularly sensitive to the values of these parameters, we have not performed extensive studies to investigate the influence of each parameter on the final result. Thus, it may be that the choice of parameters used in the current work is not ideal. This is particularly important

when dealing with experimental setups significantly different from the ones used in the current study. In such cases the users are encouraged to experiment with these parameter choices.

Finally, the method described here assigns to each protein a p value quantifying the chance that its PFMs were obtained by chance. A Bayesian framework including explicit modeling of correct and incorrect identifications may prove to be a more powerful technique, and is one of the direction in which we plan to extend the current work.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by a grant from the Swedish Research Council (to L.K.) and was funded in part by the US National Science Foundation (MRI grant No. 0923536), the American Recovery and Reinvestment Act (ARRA) funds through grant number R01 HG005805, the National Institute of General Medical Sciences 2P50 GM076547/ Center for Systems Biology and the Luxembourg Centre for Systems Biomedicine and the University of Luxembourg. The authors would like to thank William Stafford Noble, University of Washington for helpful discussions, and Chung-Ying (Alan) Huang for supplying Du145 protein digests.

References

1. Hunt DF, Yates JR 3rd, Shabanowitz J, Winston S, Hauer CR. Protein sequencing by tandem mass spectrometry. *Proceedings of the National Academy of Sciences*. 1986; 83:6233–6237.
2. McCormack A, Schieltz D, Goode B, Yang S, Barnes G, Drubin D, Yates JR 3rd. Direct analysis and identification of proteins in mixtures by LC/MS/MS and database searching at the low-femtomole level. *Analytical Chemistry*. 1997; 69:767–776. [PubMed: 9043199]
3. Wolters D, M.P. W, Yates JR 3rd. An automated multidimensional protein identification technology for shotgun proteomics. *Analytical Chemistry*. 2001; 73:5683–5690. [PubMed: 11774908]
4. Michalski A, Cox J, Mann M. More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. *Journal of Proteome Research*. 2011; 10:1785–1793. [PubMed: 21309581]
5. Second T, Blethrow J, Schwartz J, Merrihew G, MacCoss M, Swaney D, Russell J, Coon J, Zabrouskov V. Dual-pressure linear ion trap mass spectrometer improving the analysis of complex protein mixtures. *Analytical Chemistry*. 2009; 81:7757–7765. [PubMed: 19689114]
6. Olsen JV, Schwartz JC, Griep-Raming J, Nielsen ML, Damoc E, Denisov E, Lange O, Remes P, Taylor D, Splendore M, Wouters ER, Senko M, Makarov A, Mann M, Horning S. A Dual Pressure Linear Ion Trap Orbitrap Instrument with Very High Sequencing Speed. *Molecular & Cellular Proteomics*. 2009; 8:2759–2769. [PubMed: 19828875]
7. Motoyama A, Venable JD, Ruse CI, Yates JR 3rd. Automated Ultra-High-Pressure Multidimensional Protein Identification Technology (UHP-MudPIT) for Improved Peptide Identification of Proteomic Samples. *Analytical Chemistry*. 2006; 78:5109–5118. [PubMed: 16841936]
8. Iwasaki M, Miwa S, Ikegami T, Tomita M, Tanaka N, Ishihama Y. One-Dimensional Capillary Liquid Chromatographic Separation Coupled with Tandem Mass Spectrometry Unveils the *Escherichia coli* Proteome on a Microarray Scale. *Analytical Chemistry*. 2010; 82:2616–2620. [PubMed: 20222674]
9. Hubner NC, Ren S, Mann M. Peptide separation with immobilized pI strips is an attractive alternative to in-gel protein digestion for proteome analysis. *PROTEOMICS*. 2008; 8:4862–4872. [PubMed: 19003865]
10. Pappin D, Hojrup P, Bleasby A. Rapid identification of proteins by peptide-mass fingerprinting. *Current Biology*. 1993; 3:327–332. [PubMed: 15335725]

11. Henzel WJ, Billeci TM, Stults JT, Wong SC, Grimley C, Watanabe C. Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proceedings of the National Academy of Sciences*. 1993; 90:5011–5015.
12. Mann M, Højrup P, Roepstorff P. Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biological mass spectrometry*. 1993; 22:338–345. [PubMed: 8329463]
13. James P, Quadroni M, Carafoli E, Gonnet G. Protein Identification by Mass Profile Fingerprinting. *Biochemical and Biophysical Research Communications*. 1993; 195:58–64. [PubMed: 8363627]
14. Yates J, Speicher S, Griffin P, Hunkapiller T. Peptide Mass Maps: a Highly Informative Approach to Protein Identification. *Analytical Biochemistry*. 1993; 214:397–408. [PubMed: 8109726]
15. Conrads TP, Anderson GA, Veenstra TD, Pasa-Tolic L, Smith RD. Utility of accurate mass tags for proteome-wide protein identification. *Analytical Chemistry*. 2000; 72:3349–3354. [PubMed: 10939410]
16. Lu B, Motoyama A, Ruse C, Venable J, Yates JR 3rd. Improving protein identification sensitivity by combining MS and MS/MS information for shotgun proteomics using LTQ-Orbitrap high mass accuracy data. *Analytical Chemistry*. 2008; 80:2018–2025. [PubMed: 18275164]
17. Palmblad M, Ramström M, Bailey CG, McCutchen-Maloney SL, Bergquist J, Zeller LC. Protein identification by liquid chromatography–mass spectrometry using retention time prediction. *Journal of Chromatography B*. 2004; 803:131–135.
18. Bochet P, Rügheimer F, Guina T, Brooks P, Goodlett D, Clote P, Schwikowski B. Fragmentation-free LC-MS can identify hundreds of proteins. *PROTEOMICS*. 2011; 11:22–32. [PubMed: 21182191]
19. Hoopmann MR, Finney GL, MacCoss MJ. High-Speed Data Reduction, Feature Detection, and MS/MS Spectrum Quality Assessment of Shotgun Proteomics Data Sets Using High-Resolution Mass Spectrometry. *Analytical Chemistry*. 2007; 79:5620–5632. [PubMed: 17580982]
20. Identification of peptide features in precursor spectra using Hardklör and Krönik, author = Hoopmann, Michael R. and MacCoss, Michael J. and Moritz, Robert L., publisher = John Wiley & Sons, Inc., isbn = 9780471250951, doi = 10.1002/0471250953.bi1318s37, booktitle = *Current Protocols in Bioinformatics*, year = 2002,.
21. Park CY, Klammer AA, Käll L, MacCoss MJ, Noble WS. Rapid and Accurate Peptide Identification from Tandem Mass Spectra. *Journal of Proteome Research*. 2008; 7:3022–3027. [PubMed: 18505281]
22. Hsieh E, Hoopmann M, MacLean B, MacCoss MJ. Comparison of Database Search Strategies for High Precursor Mass Accuracy MS/MS Data. *Journal of Proteome Research*. 2010; 9:1138–1143. [PubMed: 19938873]
23. Käll L, Canterbury JD, Weston J, Noble WS, MacCoss MJ. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods*. 2007; 4:923–925. [PubMed: 17952086]
24. Serang O, MacCoss MJ, Noble WS. Efficient Marginalization to Compute Protein Posterior Probabilities from Shotgun Mass Spectrometry Data. *Journal of Proteome Research*. 2010; 9:5346–5357. [PubMed: 20712337]
25. Moruz L, Tomazela D, Käll L. Training, Selection, and Robust Calibration of Retention Time Models for Targeted Proteomics. *Journal of Proteome Research*. 2010; 9:5209–5216. [PubMed: 20735070]
26. Käll L, Storey JD, Noble WS. qvalue: non-parametric estimation of q-values and posterior error probabilities. *Bioinformatics*. 2009; 25:964–966. [PubMed: 19193729]
27. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*. 2002; 30:207–210. [PubMed: 11752295]
28. Prensner JR, et al. Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. *Nature Biotechnology*. 2011; 29:742–749.

29. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*. 2010; 28:511–515.
30. International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome. *Nature*. 2001; 409:860–921. [PubMed: 11237011]
31. Fujita PA, et al. The UCSC Genome Browser database: update 2011. *Nucleic Acids Research*. 2011; 39:D876–D882. [PubMed: 20959295]
32. Starmans MH, Chu KC, Haider S, Nguyen F, Seigneure R, Magagnin MG, Koritzinsky M, Kasprzyk A, Boutros PC, Wouters BG, Lambin P. The prognostic value of temporal in vitro and in vivo derived hypoxia gene-expression signatures in breast cancer. *Radiotherapy and Oncology*. 2012; 102:436–443. [PubMed: 22356756]
33. Käll L, Storey JD, Noble WS. Non-parametric estimation of posterior error probabilities associated with peptides identified by tandem mass spectrometry. *Bioinformatics*. 2008; 24:i42–i48. [PubMed: 18689838]
34. Gillet LC, Navarro P, Tate S, Röst H, Selevsek N, Reiter L, Bonner R, Aebersold R. Targeted Data Extraction of the MS/MS Spectra Generated by Data-independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis. *Molecular & Cellular Proteomics*. 2012;11.
35. Moruz L, Staes A, Foster J, Hatzou M, Timmerman E, Martens L, Käll L. Chromatographic retention time prediction for post-translationally modified peptides. *PROTEOMICS*. 2012; 12:1151–1159. [PubMed: 22577017]
36. Deutsch EW, Lam H, Aebersold R. Data analysis and bioinformatics tools for tandem mass spectrometry in proteomics. *Physiological Genomics*. 2008; 33:18–25. [PubMed: 18212004]

Figure 1. Analogy between assigning peptide confidence using fragmentation spectra and calculating protein probabilities using the MS1-features

When inferring peptides from the fragmentation spectra (panel A), we compare the theoretical spectrum of a peptide with the observed spectrum, and assign a peptide probability reflecting the quality of this match. Similarly, we can calculate protein probabilities (panel B) by comparing the theoretical peptides of a protein with the observed MS1-features.

Figure 2. PFM-derived protein p value distribution for yeast-01

We assigned for each protein a p value describing the probability that an incorrect protein would get a similar, or more extreme configuration of peptides matched to MS1-features. In (A) we display the histogram of the p values obtained for incorrect proteins composed of random peptides, while in (B) we display the p values for the proteins present in the yeast database.

Figure 3. The *in silico* AMT workflow

Our approach includes two complementary workflows. The first one, depicted on the left-side, includes the typical steps of a shotgun experiment, where we employ the fragmentation spectra to derive protein probabilities. In a second workflow, we map the peptides that did not match a fragmentation spectrum to the MS1-features, and use the resultant peptide-feature matches to derive a new set of protein probabilities. Finally, we combine the two statistics to derive improved protein-level confidence measures.

Figure 4. The PFM level information increases the number of protein identifications

In (A) we display the number of proteins as a function of the PSM-derived q values (dashed line), and combined PFM+PSM-derived q value (solid line) for the yeast-01 dataset. In (B) and (C) we show the overlap between the yeast proteins identified at $q < 0.01$ when using the *in silico* AMT workflow, and when using only the PSMs, respectively.

- (A) **>P26583**
 MGKGDPNKPGRGKMSSYAFFVQTCREEHKKK([HPDSSVNFAEFSK])KCSERWKTMSAKEKSKFEDMAKSDKA
 RYDREMKNYVPPKGDKKGKKK([DPNAPK])(RPPSAFFLCSEHRPK)IK([SEHPGLSIGDTAK])K
 [LGEMWSEQSAK]DKQPYEQKAAKLKEYEKDIAAYRAKGKSEAGKKGPRPTGSKKKNEPEDEEEEEEE
 DEEEEEDEEE
- (B) **>P48739**
 MVLIKEFR([VVLPCSVQEQVVGQLYSVAEASK])(NETGGGEGIEVLK)(NEPYEK)DGEKGQYTHKIYHLKSKVP
 AFVR([MIAPEGSLVFHEK])AWNAYPYCRTIVTNEYMKDDFFIK(IETWHKPDLTENVHGLDPNTWK)TVEI
 VHIDIADRSQVEPADYK([ADEDPALFQSVK])TKR([GPLGNWK])(ELANSPDCPQMCAYK)LVTIKFKWWGLQ
 SKVENFIQKQEKRIFTNFHRQLFCWIDKWIDLTMEDIRMEDETQKELETMRKRGSVRGTSADV
- (C) **>O95571**
 MAEAVLRVARQLSQRGSGAPILLR([QMFEPVSCFTTYLLGDR])ESREAVLIDPVLETAPRDAQLIKELGL
 R([LLYAVNTHCHADHITGSGLLR])(SLLPGCQSVISR)(LSGAQADLHIEDGDSIR)FGRFALETR
 (ASPGHTPGCVTFVLNDHSMFTGDALLIR)GCGR([TDFQQGCAK])TLYHSVHEK
 ([IFTLPGDCLYPAHDYHGFTVSTVEER])TLNPRLTLSCEEFVK([IMGNLNLPKQQIDFAVPANMR])CGVQTPTA

Figure 5. Example of confidently identified proteins after including the PFMs

The PSM-derived q values for the three proteins displayed in (A), (B), and (C) were 0.5×10^{-3} , 0.04 and 0.75, respectively. After including the PFMs, their q values became 0.2×10^{-3} , 0.5×10^{-4} and 0.7×10^{-7} , respectively. The peptides with PSM-derived q values below 0.01 are displayed between curly brackets in green color and those with PSM-derived q values above 0.01 are displayed in square brackets in orange. The peptides matched to MS1-features are given in blue between round brackets, and those that are both matched to a fragmentation spectrum (PEP= 1.0) and an MS1-feature are shown in magenta between round and square brackets.

Figure 6. Comparison of the transcript levels of PSM-derived and PSM+PFM-additional protein identifications

We have processed a publicly available dataset corresponding to mRNA profiling via high-throughput sequencing for the DU145 cells. In green we display the estimated transcript levels of the proteins confidently identified from the fragmentation spectra, and in blue the transcript levels of the additional proteins obtained with the *in silico* AMT method. The abundance levels are expressed as Fragments Per Kilobase of transcript per Million mapped reads (FPKM), and for a better visualization are given on a logarithmic scale. As a consequence, the proteins that had estimated transcript levels of 0.0 were removed (25 proteins for the PSM-proteins and 1 for the PFM-proteins). The gray line illustrates the transcript levels of all the genes, and was generated by kernel density estimation with the bandwidth set using Silverman's rule of thumb, and the values normalized to match the other two distributions. The figure corresponds to the human-01 dataset, and Supplementary Figure 12 displays similar representations for the other two human runs.

Figure 7. Coverage comparisons of proteins and PSMs

For the yeast-01 dataset, we considered all the confident PSMs with $q < 0.01$. We then selected the subset of yeast proteins that were inferred from these PSMs, and sorted them by their number of constituent PSMs in descending order. Next, we cumulatively counted the number of included PSMs when traversing the list of proteins. As an example, a point (x,y) on the graph would indicate that the fraction x of the proteins, cover a fraction y of the PSMs.

Table 1**Datasets**

Columns 2-4 display the number of peptides, proteins and MS1-features for the six datasets investigated. Column 5 gives the average $\bar{\delta}_m$ and the standard deviation σ_m of the mass errors for the peptides confidently identified from the fragmentation spectra. Column 6 gives similar statistics for the retention time errors ($\bar{\delta}_t$ and σ_t). The last two columns give the threshold used for the combined mass and retention time error, and the total number of peptide-feature matches (PFM) of each dataset.

Dataset	Peptides ($q < 0.01$)	Proteins ($q < 0.01$)	MS1 features	Avg(sd) mass error [ppm]	Avg(sd) RT error [min]	Combined error threshold r	PFMs
yeast-01	3,908	996	45,118	-0.0(0.9)	0.3(5.4)	1.5	7,404
yeast-02	3,643	972	46,185	-0.1(0.9)	0.4(4.6)	1.6	6,845
yeast-03	3,706	967	47,401	0.0(0.9)	0.2(4.9)	1.6	7,459
human-01	6,801	1,614	36,066	0.1(0.8)	-0.0(5.7)	1.6	27,428
human-02	6,687	1,622	37,633	0.1(0.8)	-0.5(5.7)	1.6	28,524
human-03	6,672	1,714	38,749	0.0(0.8)	-0.3(5.6)	1.6	27,098

Table 2
Number of proteins after combining the PFM- and PSM-derived protein probabilities

For each of the datasets investigated, we display the number of proteins at $q < 0.01$ inferred using the fragmentation spectra (second column), as well as the number of unique peptides matched to a fragmentation spectra (third column) and the number of unique peptides matched to MS1-features (fourth column) mapping to these proteins. Similarly, the fifth column gives the number of proteins at $q < 0.01$ after incorporating the peptide-feature matches, while the following two columns give the number of peptides matched to fragmentation spectra and MS1-features mapping to these proteins. The last two columns give the number and percentage of additional proteins obtained when including the peptide-feature matches (PFM). Here PSM stands for peptide-spectrum match.

Dataset	PSM proteins			PSM+PFM proteins				
	#proteins	PSM peptides ^a	PFM peptides ^b	#proteins	PSM peptides ^a	PFM peptides ^b	Additional proteins	Increase
yeast-01	996	5,370	1,231	1,054	5,481	1,438	58	5.8%
yeast-02	972	5,291	1,159	1,044	5,417	1,391	72	7.4%
yeast-03	967	5,170	1,198	1,031	5,276	1,415	64	6.6%
human-01	1,614	9,736	3,198	1,704	9,878	3,580	90	5.6%
human-02	1,622	9,752	3,362	1,715	9,896	3,821	93	5.7%
human-03	1,714	9,255	3,437	1,806	9,370	3,937	92	5.4%

^aUnique peptides matched to fragmentation spectra with posterior error probability < 1.0.

^bUnique peptides matched to MS1-features.