research articles Journal of proteome research

Peptide Sequence Confidence in Accurate Mass and Time Analysis and Its Use in Complex Proteomics Experiments

Damon May, Yan Liu, Wendy Law, Matt Fitzgibbon, Hong Wang, Samir Hanash, and Martin McIntosh*

Molecular Diagnostics Program, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109

Received June 18, 2008

We present new algorithms and a software implementation for assigning confidence to peptide sequence assignments obtained through classic accurate mass and retention time (AMT) matching techniques, as well as methods for integrating these assignments with standard proteomics workflows. The algorithms are intended to increase the number of peptides and proteins identified (and, when applicable, quantitated by isotopic labeling) among related proteomics experiments that use high-resolution mass spectrometry instrumentation. The motivations for our extensions include the need to exploit high-resolution data to support highly complex proteomics experiments, especially those involving extensive off-line fractionation, to which recent label-free workflows might not easily generalize.

Keywords: AMT • msInspect • LC-MS • IPAS

Introduction

High-resolution mass spectrometry (MS), along with tandem MS, dramatically increases the precision and volume of data that can be captured from proteomics experiments compared with tandem MS using low-resolution instruments. In particular, the high-resolution instruments provide a more complete census of precursor ions observable in a protein mixture. Several related approaches have been recently developed to exploit these data (see Veltri et al.[1] and Mueller et al.[2] for recent reviews) that rely on direct chromatographic alignment and emphasize the use of these data for label-free quantitative proteomic analysis. Here, we instead focus on the use of high-resolution LC-MS data in more traditional experiments which use isotopic labeling for quantitative comparisons and which also involve extensive fractionation of peptides and proteins.

All approaches that exploit high-resolution data begin with the identification of peptide signatures, including location of monoisotopic masses and retention times and computation of ion intensities. Approaches to downstream processing of these discovered peptide locations diverge. Some recent methods are based on associating ions across multiple experiments by direct chromatographic alignment. However, the earlier use of high-resolution data, pioneered by the Smith Laboratory,[3−5] uses an Accurate Mass and Time (AMT) method for comparing ions. AMT exploits the fact that each peptide's location in mass and normalized retention time (NRT) is strongly related to its chemical composition. The sequences of peptide ions can be determined by matching their AMT "tags" to the tags stored in an external peptide database derived from MS/MS analysis. The AMT approach has been demonstrated to find more peptides in a single MS interrogation than tandem MS alone, and ion intensities may be used for quantitative comparison across or (when isotope labeling is used) within experiments.

One of the most challenging aspects of the AMT approach is matching the ions located in a single MS interrogation to a dense AMT database containing thousands of sequence entries, such as those derived from large-scale proteomics experiments. Determining the accuracy of each AMT assignment is a key component of any AMT workflow, as it allows the ability to control the overall error rate of the experiment.[4−6]

We have developed a new algorithm for determining the confidence of sequence assignments obtained through AMT methods. The algorithm extends the approach introduced by the Smith Laboratory[5] and also the approach previously implemented in msInspect/AMT.[6]

The Smith method determines match confidence in a manner directly analogous to the decoy database approach in tandem MS,[7] in which False Identification Rates (FIRs) are computed. In brief, peptide locations are matched within some distance threshold to a target AMT database and then again matched within the same threshold to a decoy AMT database that contains the same sequences but with perturbed masses (e.g., 11 Da added to the peptide mass). The FIR is computed by comparing the numbers of matches to the target and decoy AMT databases.

One disadvantage of this and all FIR approaches is that the same uncertainty measure applies to the entire group of peptides identified within the region and does not distinguish between the higher- and lower-quality assignments within the group. Our approach attempts to identify parameters for determining match accuracy dynamically and to compute a per-peptide level of confidence (a match probability), an approach analogous to that taken by PeptideProphet[8] for evaluating tandem MS measurements.

* To whom correspondence should be addressed. E-mail: mmcintos@fhcrc.org.

10.1021/pr8004502 CCC: $40.75
Published on Web 10/28/2008

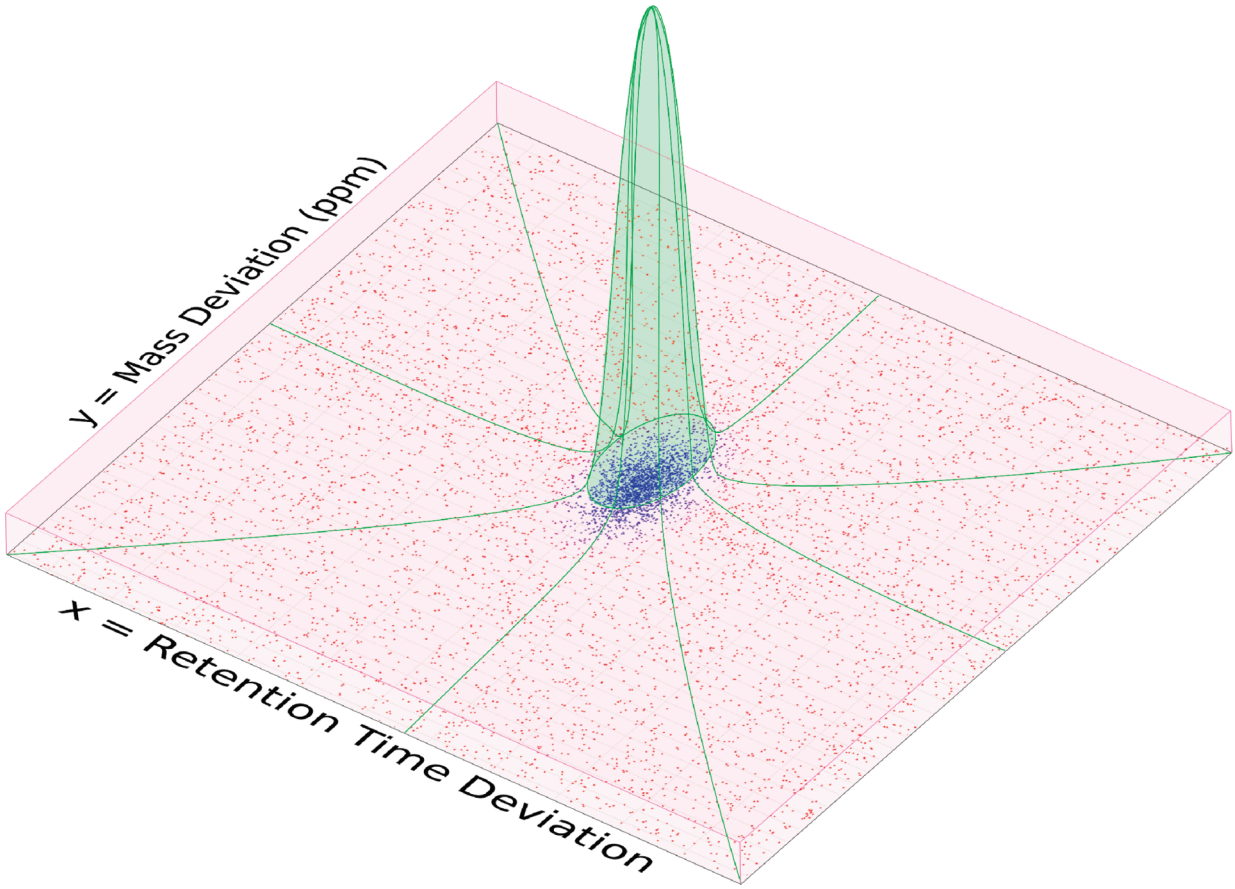

Figure 1. An illustration of the mixed distribution of AMT database matches across the two-dimensional space of Mass and NRT match error. Points indicate individual AMT matches. Red box represents the near-uniform distribution of false matches. Green region indicates the bivariate normal distribution of true matches. Coloration of individual points indicates probability as assigned by the EM algorithm: reddest points indicate $p = 0$, bluest points indicate $p = 0.96$.

We have implemented these new algorithms for making peptide assignments and assigning match confidence within the open-source msInspect/AMT software platform.[6] We also include other extensions to msInspect/AMT that support the use of these AMT-derived sequence assignments alongside sequences identified through standard tandem MS experiments. Specifically, after the assignment of peptides via AMT, msInspect/AMT automatically augments the tandem MS search results (PepXML files) to include the AMT sequences and match probabilities for use by downstream analysis tools for purposes such as protein inference (e.g., ProteinProphet[9]) or quantitation using isotope labeling (e.g., XPRESS,[10] ASAPratio[11] or Q3[12] for SILAC, ICAT, Acrylamide, or ^{18}O[13] labeling).

To demonstrate the performance of our approach, we evaluated a series of experiments using isotopically labeled human plasma, having between 88 and 96 fractions each. We show that, in this uncommonly complex experiment, accurate AMT assignments in combination with tandem MS approaches can increase the yield of confidently identified and quantitated peptides and proteins in each fraction, and in each experiment, compared with MS/MS results alone. This analysis shows that the traditional AMT workflow may be particularly useful in complex experiments having extensive fractionation, to which the more recent methods that exploit high-resolution data may not generalize well.

Experimental Procedures

We begin by presenting our algorithm for assigning confidence to sequence assignments obtained through AMT matching and also extensions to msInspect/AMT to support their use in applied experiments. We then describe the series of experiments used to evaluate these methods.

Algorithm for Assigning Peptide Sequences Using AMT and Evaluating Match Confidence. Consider a single MS interrogation in which N peptide features have been located and their retention times normalized, and which we wish to match to the locations of sequences in an AMT database (a description of the steps needed to generate an AMT database and to extract and normalize peptide features is described below). Peptides will match the AMT database elements imperfectly, and we denote these errors in mass and retention times together as $Z_i = (X_i, Y_i)$, where X_i and Y_i represent errors in mass and retention time, respectively. A density plot representing the distribution of Z for a single fraction is shown in Figure 1 showing that, as has been observed previously,[4] the distribution of errors in each dimension contains a Gaussian distribution, also mixed with an apparent uniform distribution. Here, we formally model these distributions with the hypothesis that the distribution components result from a latent (unobserved) dichotomous variable D_i representing correct ($D_i = 1$) or incorrect ($D_i = 0$) AMT assignments. Our statistical procedure will be used to estimate the latent quantities and use these estimates as the level of confidence in each AMT match. This formulation is functionally a two-dimensional version of the approach used with tandem MS identifications by PeptideProphet, in which the distribution of the null component (a Gamma distribution in PeptideProphet) is

replaced by a uniform distribution. We refer the reader to the PeptideProphet manuscript[8] for technical understanding of this approach in a context specific to proteomics and to Dempster et al.[14] for technical details of the statistical approach in general.

a. Represent Mass and Time Match Errors Using a Statistical Model. Formally, we model D_i marginally as a Bernoulli distribution with probability p (the total rate of correct assignments) and model Z for false matches ($D_i = 0$) as two independent uniform distributions over the matching tolerances t_x and t_y, with areas $A_x = [-t_x, t_x]$ and $A_y = [-t_y, t_y]$. For true matches ($D_i = 1$), we approximate Z with two independent normal distributions with mean and standard deviation μ_X, σ_X, μ_Y, and σ_Y, respectively. Figure 1, as well as analysis of many such distributions, supports each of these assumptions. Thus, formally,

$$P(Z|D) = \begin{cases} \dfrac{1}{A_x A_y} & D = 0 \\ \dfrac{1}{\sigma_x}\varphi\left(\dfrac{x-\mu_x}{\sigma_x}\right)\dfrac{1}{\sigma_y}\varphi\left(\dfrac{x-\mu_y}{\sigma_y}\right) & D = 1 \end{cases} \quad (1)$$

b. Estimate Model Parameters. We estimate the model parameters by their maximum likelihood estimates (MLE), making use of the Expectation-Maximization (EM) algorithm[14] as a device to compute them. We omit technical details of the EM algorithm because the iterative steps for our model are quite similar to those described in detail elsewhere.[10] However, in brief, in this specific mixture model framework (this is not true for all statistical models), the EM algorithm reduces to an intuitive, simple iterative procedure in which we first replace the latent data elements D_i with their expected values (denoted $\hat{d}_i$) computed as if the model parameters were known and then estimate the parameters as if the missing data elements D_i are equal to $\hat{d}_i$. The first step (the E-Step) can be written as follows:

$$\hat{d}_i = P(D = 1|Z) = \left(\frac{\hat{p}(P(Z|D = 1))}{(1 - \hat{p})(P(Z|D = 0)) + \hat{p}(P(Z|D = 1))}\right) \quad (2)$$

The second step (M-step) can be expressed as $\hat{p} = \sum \hat{d}_i / N$ and $\hat{\mu}_x = \sum w_i x_i$ and $\hat{\sigma}_x = (\sum w_i x_i^2 - \hat{\mu}_x^2)^{1/2}$ (similar expressions hold for $\hat{\mu}_y$ and $\hat{\sigma}_y$) where $w_i = \hat{d}_i / \sum \hat{d}_i$.

c. Choose Algorithm Starting Point and Evaluating Convergence. The iterative EM algorithm requires a starting point and also a method for determining convergence of the algorithm. The EM algorithm is quite robust to the specific choice of starting parameters, and so it is most convenient to begin with computationally simple approximations. Our starting point sets μ_X and μ_Y to the mean of all error values in the RT and mass dimensions, respectively, and we set σ_X and σ_Y to their standard deviations. The starting point for p is derived from the FIR approximated from a decoy AMT match using loose tolerances (as previously implemented in msInspect/AMT). To evaluate convergence, we follow standard approaches and monitor all parameters and the complete data likelihood, but we also monitor the results of the E-steps, $\hat{d}_i$, which provide our assignment probability estimates. We stop when the largest change in any assigned probability between iterations is smaller than 0.5% (for a probability of 0.9, for instance, this represents a change of 0.0045), or at minimum after 30 iterations. msInspect also provides graphs which can be used to evaluate convergence, including a plot of the model parameters and probability change estimates against iterations [see Supporting Information for example].

d. Filtering Identifications. A general filter is applied to remove the obvious errors. We remove all sequence assignments having probability less than 0.1, and those for which the second best match exceeds 0.5 or the first and second best match are within 0.1 of each other (all parameters configurable).

Integrating Matching Results into Standardized Pipeline. The algorithm provides estimates of $\hat{d}_i$, the probability of AMT assignment i being correct. msInspect/AMT adds all of the AMT sequences with their matching probabilities to the PepXML files resulting from the MS/MS search from the same interrogation, so that they may be used by all downstream analysis tools that operate on this standard file format, including ProteinProphet for protein inference and quantitation tools such as EXPRESS, Q3 or ASAPratio.

Interrogation of Plasma Using Isotope Labeling and Intact Protein Separation Prior to LC-MS/MS. Four independent, matched pairs of human serum pools were interrogated and compared with the Intact Protein Analysis System (IPAS).[12,15] In brief, for each experiment, consisting of one disease pool and one control pool, sera pools were separately depleted of the top six abundant serum proteins using a Multiaffinity Removal System (MARS) column (4.6 × 100 mm; Agilent, Wilmington, DE);[15] then, intact proteins were labeled with either heavy or light acrylamide[12] and combined prior to extensive off-line separation.[15] The separation strategy used an orthogonal two-dimensional HPLC system in which intact proteins are fractionated first on an anion exchange column and then on a reversed phase column for a total of 656 fractions. These fractions were pooled into 96 fractions (fewer fractions were collected in some experiments due to equipment malfunction), digested,[15] then interrogated using high-resolution tandem MS using an LTQ OrbiTrap XL mass spectrometer (Thermo-Finnigan) coupled with a nanoLC 2D, a two-dimensional HPLC system (Eksigent). The spectra were acquired in a data-dependent mode in m/z range of 400−1800, with selection of the five most abundant +2 or +3 ions of each MS spectrum for MS/MS analysis.

Database Search of MS/MS Data and Quantitation of Isotopically Labeled Peptide Using Tandem MS Workflow. Raw data files were converted to mzXML format using ReAdW 1.1 and Xcalibur 2.2. All mzXML files were searched using X! Tandem (2007.01.01) with an alternative scoring plugin[16] compatible with PeptideProphet. Searches were conducted against the human International Protein Index database (IPI Human v3.20) plus common contaminants. All searches used the following parameters: ±1.5 Da precursor mass error, tryptic cleavage with up to two missed cleavage sites, static modification of 71.0366 Da (light acrylamide) on cysteine, potential modifications of 74.0466 Da (^{13}C acrylamide) on cysteine, and 15.9949 Da (oxidation) on methionine. Peptide assignments were evaluated using PeptideProphet.[8]

Creation of AMT Database from MS/MS Identifications and Identification of Peptide Locations in High-Resolution Data. Peptide identifications from the LC/MS-MS database search were processed using previously described methods which place retention times on a common scale.[6,17] We included in the AMT database all peptides with PeptideProphet probability ≥0.95. Each of the 374 mzXML files were processed by msInspect to discover all LC-MS peptide locations,[18] which were filtered for quality by removing all peptides located without multiple isotopes or with a KL score exceeding 3.0 (KL is a quality score for LC-MS peptides[18]). Normalization procedures[6] were used to place their retention times on the same normalized scale as the AMT database. AMT database entries

Table 1. Summary of Peptide-Level and Protein-Level Results for Each Experiment, with MS/MS Data Alone and with MS/MS Data Combined with AMT Data[a]

		experiment-level summary				fraction-level summary			
experiment (number of fractions)	analysis approach	unique peptides	quant. peptides	unique proteins	quant. proteins	unique peptides	quant. peptides	unique proteins	quant. proteins
1 (96)	MS/MS	6128	1726	841	392	722.7	255.2	145.9	77.5
	+AMT	7357	1922)	1038	428	1165.3	343.9	213.9	97.3
	(% Increase)	(20.1%)	(11.4%	(23.4%)	(9.2%)	(61.2%)	(34.8%)	(46.6%)	(25.5%)
2 (96)	MS/MS	7230	1408	1103	329	742.4	209.3	160.0	60.9
	+AMT	8204	1603	1227	354	1045.5	253.3	218.5	71.7
	(% Increase)	(13.5%)	(13.8%)	(11.2%)	(7.6%)	(40.8%)	(21.0%)	(36.6%)	(17.7%)
3 (88)	MS/MS	5843	1401	710	270	760.6	214.9	142.3	65.9
	+AMT	7069	1632	875	314	1069.6	270.7	191.5	77.1
	(% Increase)	(21.0%)	(16.5%)	(23.2%)	(16.2%)	(40.6%)	(26.0%)	(34.6%)	(17.0%)
4 (94)	MS/MS	7400	1774	1044	346	996.6	298.3	177.4	83.4
	+ AMT	8815	2003	1335	403	1112.9	419.9	268.1	108.9
	(% Increase)	(19.1%)	(12.9%)	(27.9%)	(16.5%)	(11.7%)	(40.8%)	(51.1%)	(30.6%)
Mean (93.5)	**MS/MS**	**6650.3**	**1577.3**	924.5	334.3	805.6	**244.4**	**156.4**	**69.2**
	+ AMT	**7861.3**	**1790.0**	**1118.8**	**374.5**	**1098.3**	**322.0**	**223.0**	**88.8**
	(% increase)	**(18.2%)**	**(13.5%)**	**(21.0%)**	**(12.0%)**	**(36.3%)**	**(31.7%)**	**(42.6%)**	**(28.3%)**

[a] All peptide counts are with PeptideProphet or AMT probability ≥0.9. All protein counts are of protein groups with ProteinProphet probability ≥0.9. Fraction-level numbers are averages over all fractions in each experiment. Counts of unique and quantified proteins per fraction are counts of protein groups with any high-quality peptide evidence (with isotopic ratios, for quantified summary) for the group in the fraction.

were duplicated to accommodate both light and heavy isotopic labels. We performed a first-pass match to the AMT database with loose mass and normalized retention time tolerances (defaults 20 ppm and 0.15 NRT units), and then, masses were calibrated based on this initial match prior to matching using the EM algorithm.

Use of Mixture Model To Assign Peptide Features to AMT Database and Augment Search Results. We next assigned each peptide location to the AMT database and inferred match confidence using the EM algorithm described above. All data following the database search were processed using an Intel Xeon 5160 3 GHz processor with 16 GB of memory (only 1 GB of memory was given to msInspect/AMT). Creating the combined AMT database from 374 fractions consumed approximately 13 min. Matching of all 374 fractions to the AMT database required 181 min (29 s per fraction). Matches passing a confidence threshold (probability ≥0.1, configurable) were added as additional information into the results of a database search on the tandem MS data for the same run.

Processing Augmented PepXML File To Identify Quantitative Ratios and Infer Proteins. Next, we computed quantitative ratios between case and control samples (light and heavy labels) using Q3,[12] an algorithm specifically designed to accommodate the 3-Da mass difference for singly labeled peptides, and ratio information was added to the existing PepXML files. Finally, we performed protein inference with ProteinProphet[9] in order to determine the proteins present in the experiment, using all identified peptides with probability greater than 0.2, and combined all peptide-level quantitation information for each identified protein in order to determine abundance ratios.

Architecture and Software Availability. All methods are implemented as part of the msInspect/AMT platform, which is a cross-platform and largely written in Java, with some statistical components (e.g., EM algorithm) written in the R statistical language. All analytical tools described in this work are freely available and open source under the Apache 2.0 license. The tools and source code, with sample data sets and a tutorial on use of the software, may be downloaded at http://proteomics.fhcrc.org/CPL/amt.

Results

A total of four experiments consisting of 374 fractions were interrogated by MS/MS. The peptide and protein identifications resulting from the traditional MS/MS analysis alone and combined with the AMT results are summarized in Table 1. The experiment-level and fraction-level data are summarized in the right and left halves of the table, respectively.

We first consider the identifications from MS/MS analysis alone. In total, between 5843 and 7400 (average 6650.3; see final row of Table 1) unique peptide sequences (PeptideProphet probability ≥0.9) were identified per experiment, and between 722.7 and 996.6 unique peptide sequences (average 805.6) were identified per individual fraction. The number of unique quantified peptides (containing at least one cysteine) ranged between 1401 and 1774 per experiment, and between 209.3 and 298.3 per fraction. On the protein level (right two columns) between 710 and 1044 protein groups (average 924.5) were identified per experiment (ProteinProphet probability ≥0.9), and within each experiment, peptide evidence for between 60.9 and 83.4 protein groups was identified on average per fraction (average over all experiments 69.2). The total number of quantified proteins was between 270 and 392 (average of 334.3) per IPAS experiment.

We also characterized each protein in each experiment by the percent amino acid coverage obtained, and also the number of fractions in which it was identified. On average, peptides associated with the accession number of each individual protein were observed in 11.9 fractions in a single experiment, and the median percent of amino acid coverage for each protein was 16.19% (95% of proteins' coverage exceeds 3.74%). We report this information because, along with the goal of identifying as many proteins as possible in an experiment, another is to improve the ability to identify different protein isoforms,[15] or proteins having a different chemical compositions but which have the same accession number (e.g., modifications, cleavage products, etc. are each different chemically but have the same accession number), and amino acid coverage information is vital to this analysis.

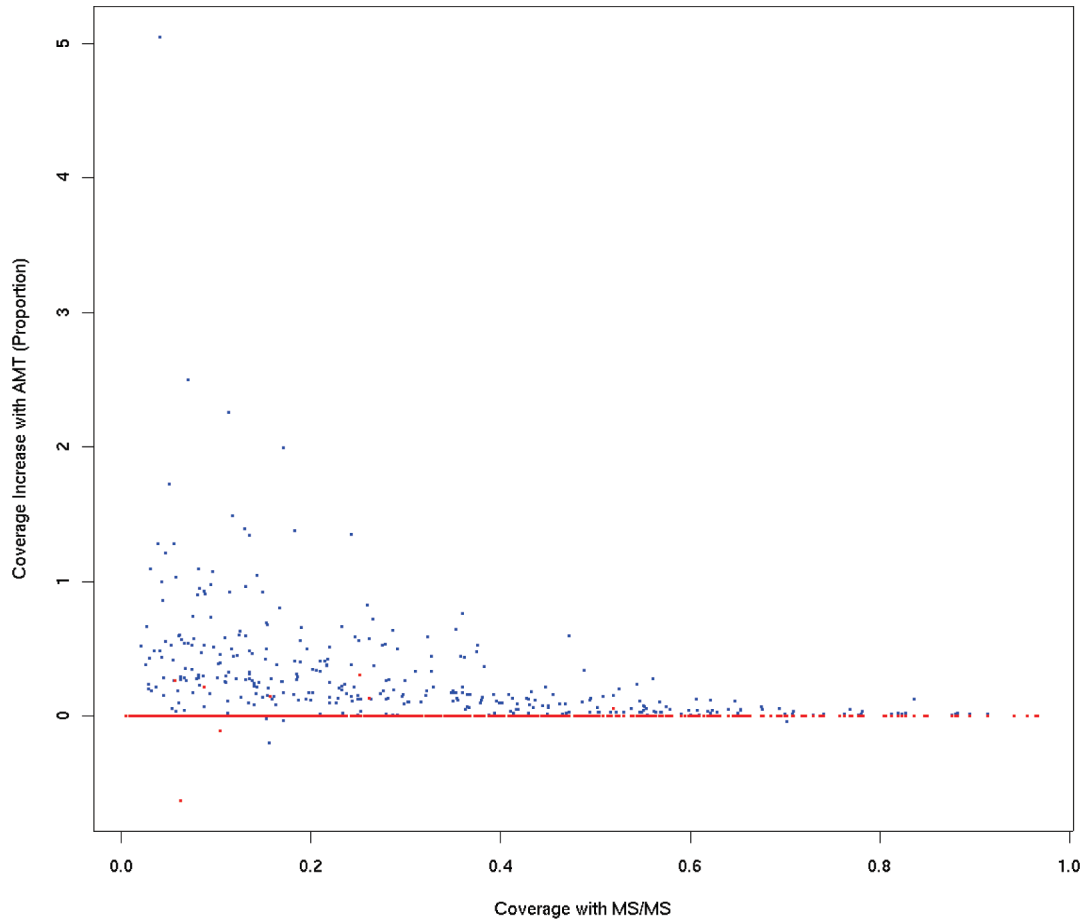

Figure 2. Fold increase in amino acid coverage of proteins with AMT peptide identifications (vertical axis) vs percent amino acid coverage using only MS/MS peptide identifications (horizontal axis). Blue points show coverage increase when matching to a target AMT database; red points show a decoy database match (only 6 proteins with increased coverage).

We next evaluated these same performance metrics using the combined MS/MS and AMT information. The increased information for all experiments is shown in Table 1. On the experiment level (left half of the table), between 7069 and 8815 (average 7861.25; see final row of Table 1) unique peptide sequences (PeptideProphet probability ≥0.9) were identified via MS/MS and AMT combined, an increase of 18.2% on average over MS/MS alone. Between 1045.5 and 1165.3 unique peptide sequences (average 1098.3) were identified per fraction, an increase of 36.3%. The number of unique quantified peptides ranged between 1603 and 2003 per experiment, an increase of 13.5%, and between 253.3 and 419.9 per fraction, an increase of 31.7%. On the protein level (right two columns), between 875 and 1335 proteins (average 1118.8) were identified with high confidence per experiment (an increase of 21.0%), and within each experiment, peptide evidence for between 191.5 and 268.1 proteins was identified on average per fraction (average over all experiments 223.0, an increase of 42.6%). The total number of quantified proteins is between 314 and 428 (average of 374.5) per IPAS experiment, an increase of 12.0%.

Every fraction found quantified peptides and proteins that were not quantified using traditional MS/MS-based approaches, and over all experiments, an average of 1113.0 peptides per experiment were quantified in fractions in which they had not been quantified by MS/MS alone (data not shown). From among all IPAS experiments, a total of 4621.25 unique peptides (69.2% of all peptides found in via MS/MS search) were identified in at least one fraction with AMT but not (in that fraction) by standard MS/MS methods.

To interpret the gain in proteins at the experiment and fraction level, one must not only consider the number of entirely new proteins ascribed to the experiment or fraction, but also the ability to increase the explained amino acid coverage of the proteins already identified. Figure 2 demonstrates this ability graphically for proteins in a single representative fraction. The horizontal axis represents the coverage based on MS/MS alone and the vertical axis represents the fold increase in coverage based on the combined analysis. Overall, of the proteins identified by high quality with MS/MS, 20% find an increase in explained amino acid sequence coverage, with a median improvement of 18% per protein.

The complementarity of AMT and MS/MS identifications, which governs the amount of increase in peptide coverage that AMT identifications provide in an MS/MS experiment, is illustrated in Figure 3. Each point represents the MS/MS (horizontal axis) and AMT (vertical axis) match probabilities for a peptide assigned by both methods in the same fraction. Region A denotes the peptides that are found by both methods with high quality (probability > 0.9; 41% of peptides fall in region A). The peptides falling in Region B are those peptides with low MS/MS PeptideProphet score in a fraction but which are confidently identified using AMT; in this experiment, 10% of peptides fall in this range. The peptides falling in Region D are those peptides with low AMT probability score in a fraction

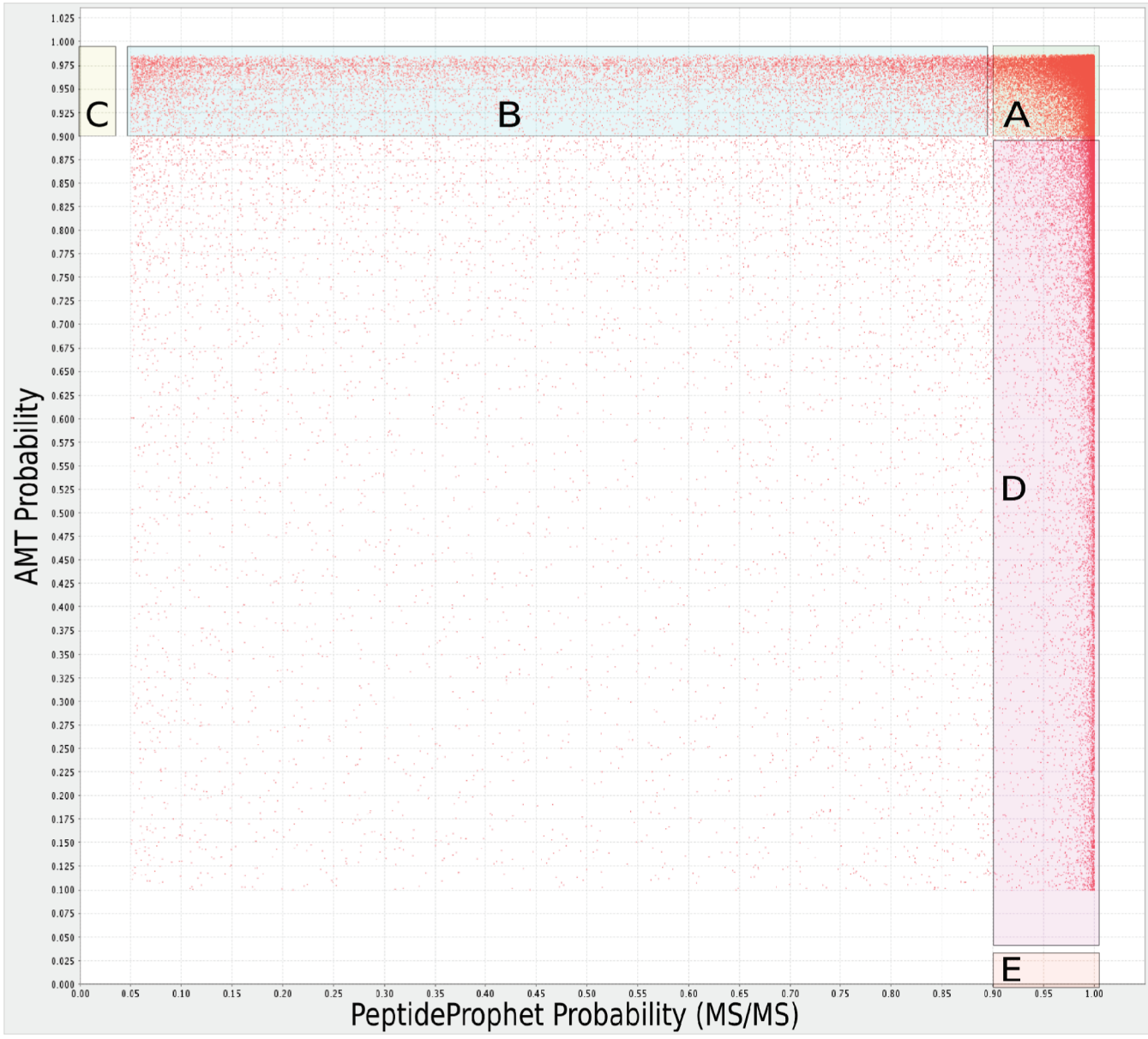

Figure 3. MS/MS database search probability (horizontal axis) vs AMT match probability for the same peptide (vertical axis). Region A (41% of peptides): high probability in both MS/MS and AMT. Region B (10%): high-probability in AMT but low in MS/MS. Region C (14%): high-probability matches unique to AMT. Region D (5%): high-probability in MS/MS but low in AMT. Region E (27%): matches unique to MS/MS.

but which are confidently identified using MS/MS; in this experiment, 5% of peptides fall in this range. Not visible in this Figure are the peptides identified by AMT (Region C, 14% of peptides) or by MS/MS (Region E, 27% of peptides), but not by both. Overall, in this experiment, the AMT approach can thus improve the number of peptides confidently identified per fraction by 14% + 10% = 25% compared with using MS/MS alone.

Evaluating the Accuracy of Matching. The results above show the ability to increase coverage and identifications based on AMT matching with the new algorithm. We used several approaches to demonstrate the overall accuracy of our matching algorithm.

The rate of agreement between MS/MS and AMT sequences identified in each fraction, as shown in Figure 3, provides a direct demonstration that the AMT matching is of high quality. We also evaluated the rates at which the AMT assignments and high-quality MS/MS assignments for the same ion disagree. We associated MS/MS identifications with LC-MS peptide features in the same mzXML file if they fell within 5 ppm and 20 s of each other and matched uniquely. Of those peptide features that matched the AMT database with probability ≥0.9, only 0.26% (a rate of 0.0026) disagreed with the sequence assigned by MS/MS. These rates suggest that the matching algorithms rarely create discordance between AMT and MS/MS identifications.

We also established that the increase in percentage of amino acid coverage one should expect by chance is far below that shown in Figure 2, by matching that same experiment to a decoy AMT database. The points in red show the results of this analysis. Compared with 20% of the proteins increasing their coverage in the target AMT database, only six proteins (less than 0.01%) found an increase when using the decoy database.

We also evaluated the fit of the parametric model using a quantile-quantile plot of the estimated mixed distribution

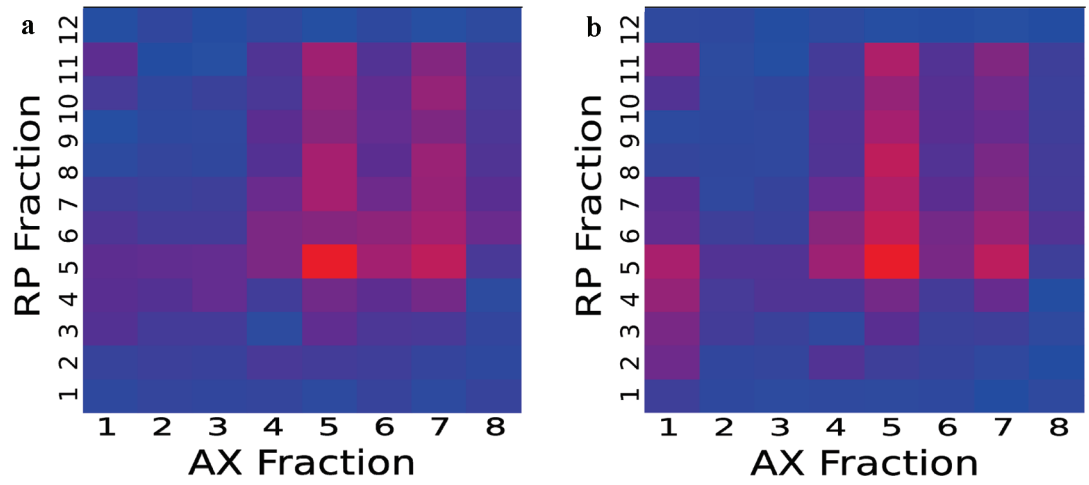

Figure 4. Heatmaps describing the distribution of peptide identifications throughout the AX (horizontal axis) and RP (vertical axis) dimensions of a fractionated experiment. Red indicates many IDs, blue indicates few. Identifications charted are those peptides confidently identified via MS/MS in fraction (5,5), the reddest fraction in both charts. (a) Peptides confidently identified via MS/MS. (b) Peptides confidently matched via AMT.

against the density of the actual match data using the automated graphing functions of msInspect (as described above; see Supporting Information for example) and found an overall high-quality agreement between the estimated parametric model and the empirical behavior of the data, suggesting that our parametric assumptions are reasonable.

Finally, we compared the spatial distribution of peptides across the two dimensions of separation within a single experiment to determine whether the MS/MS and AMT identifications were overall concordant. The similarity of the spatial distribution is evidence for the accuracy of the method because the fraction information was not used as part of the matching algorithm, and so is an independent confirmation of the accuracy of the model. To compare the spatial behavior, we selected all peptide sequences found in a selected fraction (the "origin") and then recorded the number of these peptides found in all other fractions separately by either MS/MS or AMT methods. Figure 4a shows a representative distribution of these counts across all fractions for MS/MS data, and Figure 4b shows the distribution for AMT matches; the two charts reveal a high degree of spatial association of their identified sequences.

One should also expect a high degree of correlation between the quantitative ratios derived by the Q3 algorithm based on MS/MS data and AMT data only if a high degree of accuracy can be obtained in our matching, because our matching algorithm does not make use of ion intensity. However, one should not expect identical quantitation by MS/MS and AMT for computed ion intensities because each method uses a different starting point for peptide abundance detection. The correlation coefficient between log AMT ratios (based on *de novo* discovery of peptides and quantitation) and log MS/MS ratios (based on MS/MS driven quantitation) was 0.95 (see Figure 5); 95% of the ratios (on the raw scale) differ between AMT and MS/MS by less than 15%, and 86% of the ratios peptides differ by less than 5% between the two methods. These differences compare quite favorably with, for instance, the agreement expected between different MS/MS-based quantitation methods (e.g., Q3, Xpress, ASAPRatio).

Discussion

We presented (1) a new algorithm for determining the probability of correct AMT sequence assignments, (2) its implementation in a workflow that allows the incorporation of the results into a traditional tandem MS pipeline, and (3) an example using this workflow with a set of experiments of uncommon complexity. This work leads us to conclude that it is feasible to borrow strength across a large number of experiments and across fractions within an experiment to increase the peptide and protein identifications and to increase the amino acid coverage of proteins identified by MS/MS methods alone. Since our implementation makes use of standard file formats for MS/MS data processing (PepXML), it is convenient to augment an existing MS/MS-based proteomics workflow to gain the benefits of AMT data.

The AMT approach used here follows the original formulation advocated by Smith et al., which can be seen as a means to integrate high resolution data into proteomics experiments. As with the Smith formulation, our model of the distribution of AMT matching error uses both the mass component and the NRT component of the error. This is far more effective than using only one component or the other alone. The actual benefit of using both mass and time will depend on the density of the peptide features in a single interrogation, as well as the density of the AMT database being matched. For our specific example here, using a model based only on mass, or only on NRT, to make AMT assignments would result in a roughly 4-fold increase in the number of ambiguous matches (matching assignments to more than one peptide).

Our work has focused on quantitation of peptides and proteins only for isotopically labeled experiments. This may be contrasted to the more recently developed platforms that use high-resolution data[19,20] and that emphasize label-free quantitative approaches. Those recently developed platforms circumvent the need to create an explicit AMT database and instead rely on direct chromatographic alignment of peptide locations between related series of experiments. In the AMT approach, peptide locations are associated between experiments only if they match the same entry in an external database. Each of these two approaches has advantages and disadvantages. An advantage of the direct chromatographic alignment approaches is that peptides that may never have been sequenced successfully are accessible for quantitative comparison between experiments, whereas the classic AMT

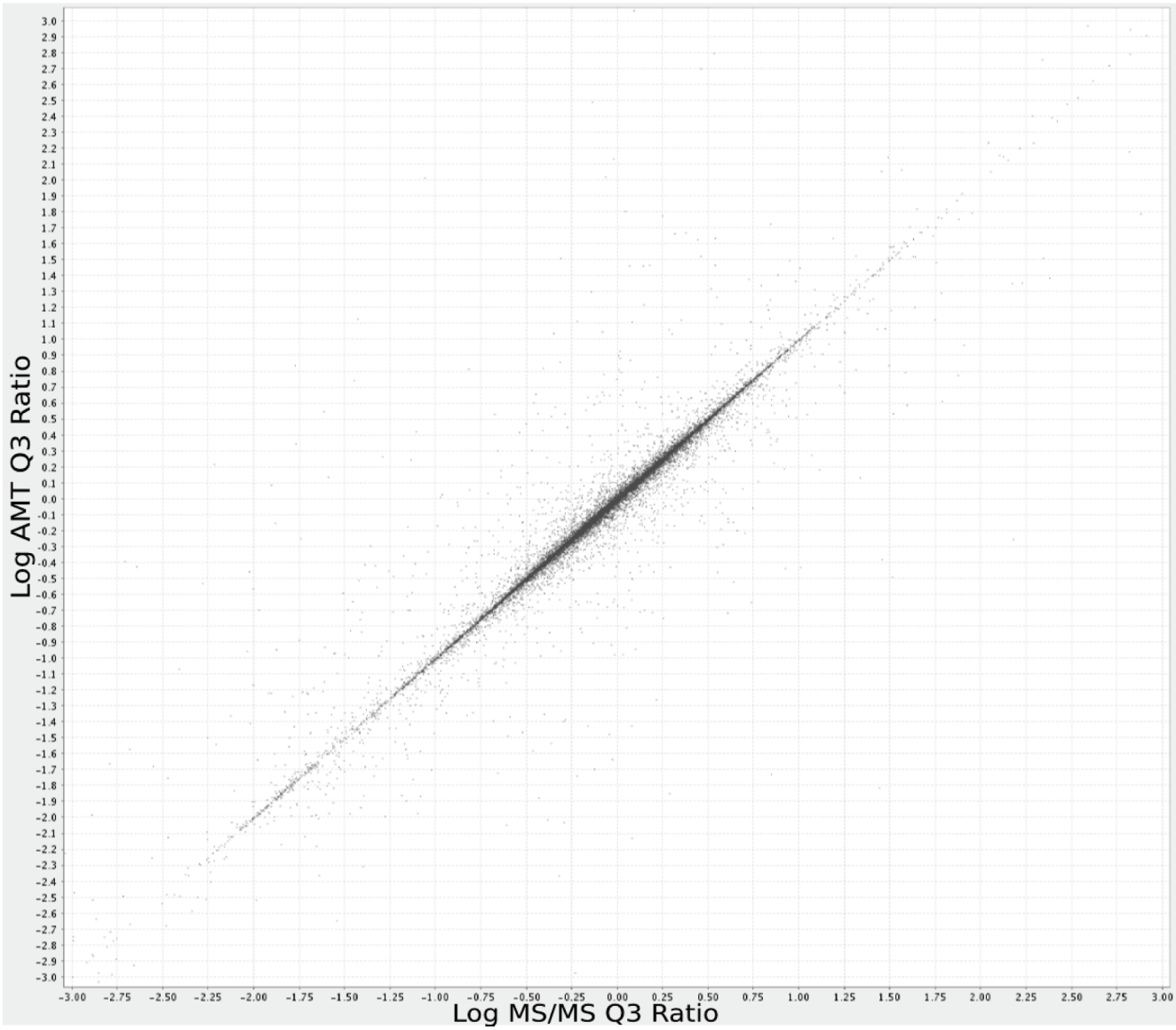

Figure 5. Comparison of median per-charge-state peptide labeled isotope ratios calculated by the Q3 algorithm, based on high-quality LC-MS/MS database search results (horizontal axis) and high-quality AMT matches (vertical axis). Correlation coefficient is 0.95.

approach requires the ions to have been selected for CID and sequenced with high confidence in some experiment.

However, an advantage of the AMT approach, combined with isotopic labeling, may be its suitability for evaluating complex experiments, such as those requiring off-line separation. There are several outstanding problems that still need to be solved before direct chromatographic alignment approaches may be used for these more complex workflows. For example, with fractionation, consider a series of n samples having k fractions each. Because individual peptides are likely to occur in multiple fractions (especially with intact protein separation[15]), naïve, direct chromatoghraphic alignment of all pairs of fractions could require roughly $(nk)^2$ alignments (precisely $nk(nk - 1)/2$), each with some propensity to admit and propagate errors. The classic AMT approach requires only nk alignments. Moreover, another consequence of peptides existing in multiple fractions is the difficulty in defining peptide intensity for peptides that migrate across one or more fractions, a problem automatically accounted for in experiments using isotopic labeling. Until those computational issues are resolved, the AMT approach using isotopic labeling and the method we describe here allow the use of information contained in high resolution instruments in complex proteomics workflows requiring separations.

Acknowledgment. Funding provided by National Cancer Institute grants U01 CA111273 and R01 CA107209, by Department of Defense grant W81XWH-06-1-0100, and by the Canary Foundation.

Supporting Information Available: A tutorial for msInspect/AMT platform, with sample charts; a version of the msInspect/AMT software; a tutorial with sample data for demonstration of AMT matching and confidence estimation; additional charts referred to in this manuscript. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Veltri, P. *Briefings Bioinf.* **2008**, *9*, 144–155.
(2) Mueller, L. N.; Brusniak, M.; Mani, D. R.; Aebersold, R. *J. Proteome Res.* **2007**, *7*, 51–61.

(3) Smith, R. D.; erson, G. A.; Lipton, M. S.; Pasa-Tolic, L.; Shen, Y.; Conrads, T. P.; Veenstra, T. D.; Udseth, H. R. *Proteomics* **2002**, *2*, 513–23.
(4) Norbeck, A. D.; Monroe, M. E.; Adkins, J. N.; erson, K. K.; Daly, D. S.; Smith, R. D. *J. Am. Soc. Mass Spectrom.* **2005**, *16*, 1239–49.
(5) Petyuk, V.; Qian, W. J.; Chin, M.; Wang, H.; Livesay, E.; Monroe, M. E.; Adkins, J.; Jaitly, N.; erson, D.; Camp, D. G., II; Smith, D. J.; Smith, R. *Genome Res.* **2007**, *17*, 328–336.
(6) May, D.; Fitzgibbon, M.; Liu, Y.; Holzman, T.; Eng, J.; Kemp, C. J.; Whiteaker, J.; Paulovich, A.; McIntosh, M. *J. Proteome Res.* **2007**, *6*, 2685–2694.
(7) Elias, J. E.; Gygi, S. P. *Nat. Methods* **2006**, *22*, 2830–2832.
(8) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2002**, *74*, 5383–5392.
(9) Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2003**, *75*, 4646–4658.
(10) Han, D. K.; Eng, J.; Zhou, H.; Aebersold, R. *Nat. Biotechnol.* **2001**, *19*, 946–951.
(11) Li, X. J.; Zhang, H.; Ranish, J. R.; Aebersold, R. *Anal. Chem.* **2003**, *75*, 6648–6657.
(12) Faca, V.; Coram, M.; Phanstiel, D.; Glukhova, V.; Zhang, Q.; Fitzgibbon, M.; McIntosh, M.; Hanash, S. *J. Proteome Res.* **2006**, *5*, 2009–2018.
(13) Stewart, I. I.; Thompson, T.; Figeys, D. *Rapid Commun. Mass Spectrom.* **2001**, *15*, 2456–2465.
(14) Dempster, A. P.; Laird, N. M.; Rubin, D. B. *J. R. Stat. Soc., Ser. B* **1977**, *39*, 1–38.
(15) Faca, V.; Pitteri, S. J.; Newcomb, L.; Glukhova, V.; Phanstiel, D.; Krasnoselsky, A.; Zhang, Q.; Struthers, J.; Wang, H.; Eng, J.; Fitzgibbon, M.; McIntosh, M.; Hanash, S. *J. Proteome Res.* **2007**, *6*, 3558–3565.
(16) MacLean, B.; Eng, J.; Beavis, R. C.; McIntosh, M. *Bioinformatics* **2006**, *22*, 2830–2832.
(17) Krokhin, O. V.; Craig, R.; Spicer, V.; Ens, W.; Standing, K. G.; Beavis, R. C.; Wilkins, J. A. *Mol. Cell. Proteomics* **2004**, *3*, 908–919.
(18) Bellew, M.; Coram, M.; Fitzgibbon, M.; Igra, M.; Randolph, T.; Wang, P.; May, D.; Eng, J.; Fang, R.; Lin, C.; Chen, J.; Goodlett, D.; Whiteaker, J.; Paulovich, A.; McIntosh, M. *Bioinformatics* **2006**, *22*, 1902–1909.
(19) Jaffe, J. D.; Mani, D. R.; Leptos, K. C.; Church, G. M.; Gillette, M. A.; Carr, S. A. *Mol. Cell. Proteomics* **2006**, *5*, 1927–1941.
(20) Wang, P.; Coram, M.; Tang, H.; Fitzgibbon, M. P.; Zhang, H.; Yi, E.; Aebersold, R.; McIntosh, M. *Biostatistics* **2006**, *8*, 357–367.

PR8004502