Impact of Amidination on Peptide Fragmentation and Identification in Shotgun Proteomics

Sujun Li,* Aditi Dabir,† Santosh A. Misal,‡ Haixu Tang,† Predrag Radivojac,*† and James P. Reilly*†‡

†School of Informatics and Computing, Indiana University, Bloomington, Indiana 47405, United States
‡Department of Chemistry, Indiana University, Bloomington, Indiana 47405, United States

ABSTRACT: Peptide amidination labeling using S-methylthioacetimidate (SMTA) is investigated in an attempt to increase the number and types of peptides that can be detected in a bottom-up proteomics experiment. This derivatization method affects the basicity of lysine residues and is shown here to significantly impact the idiosyncrasies of peptide fragmentation and peptide detectability. The unique and highly reproducible fragmentation properties of SMTA-labeled peptides, such as the strong propensity for forming b1 fragment ions, can be further exploited to modify the scoring of peptide-spectrum pairs and improve peptide identification. To this end, we have developed a supervised postprocessing algorithm to exploit these characteristics of peptides labeled by SMTA. Our experiments show that although the overall number of identifications are similar, the SMTA modification enabled the detection of 16−26% peptides not previously observed in comparable CID/HCD tandem mass spectrometry experiments without SMTA labeling.

KEYWORDS: mass spectrometry, proteomics, chemical modification, peptide fragmentation, amidination, peptide identification, fragment ions, logistic regression

INTRODUCTION

Over the past decade, mass spectrometry has become the dominant platform in proteome research.1−5 Advances in technology and algorithm development have recently enabled deep proteome coverage of several model organisms with the current draft of the human proteome reaching 92%.6 Shotgun deep proteome coverage of several model organisms with the technology and algorithm development have recently enabled resolution,13,14 and modiﬁcations to proteins and coatings has led to improvements in chromatographic resolution of peptides.15,16 Furthermore, peptides have been chemically modiﬁed to enable N- and C-terminal fragments to be distinguishable and to modify ion fragmentation patterns.17,18 Typically, for unmodiﬁed tryptic peptides, the collision-induced dissociation (CID) spectra are dominated by b- and y-type fragment ions. The latter tend to be more abundant than the former because b ions undergo secondary fragmentation more readily than y ions, thus creating small fragments that may not be detectable with the ion trap. Multiple strategies have been developed to alter the ion fragmentation patterns to improve peptide detectability by charge-remote or charge-induced processes.19−21 The locations of modiﬁed groups in peptides can directly inﬂuence the locations of protons and thus inﬂuence the fragment ion patterns; in addition, altering fragmentation by changing the basicity or acidity of residues may lead to different peptide identiﬁcations by mass spectrometry.18 Beyond the well-established database searching methods, computational researchers have incorporated this type of information into database searching and postprocessing of search results.22−25

We have previously studied labeling techniques and provided anecdotal evidence that peptide amidination leads to changes in peptide basicity thereby affecting their ionizability26 and fragmentation.27,28 This is because N-terminal amidination leads to the formation of cyclic intermediates that enable peptide bond cleavage, stabilizing of otherwise unobservable b1 ions and subsequent intense yN−1 ions.26−28 We have also systematically investigated peptide fragmentation and detectability.29−32 To follow up on that work, the present study focuses on quantifying the impact of amidination on peptide
fragmentation and detectability of a large number of peptides. By exploiting the idiosyncrasies of peptides containing this label, we have then developed a new data interpretation algorithm to postprocess peptide assignments from standard database search engines that ultimately lead to an improved diversity of identified peptides and better proteome coverage.

■ METHODS

Sample Preparation

*Escherichia coli* K12 MG1655 cells were grown at 37 °C in Luria Broth media for 16 h. Cells were harvested by centrifugation at 7000 rpm for 10 min using a JA10 rotor. The bacterial pellet was resuspended in lysis buffer (25 mL of Speeding buffer A, 8.75 μL of 2-mercaptoethanol, and three protease inhibitor tablets) and lysed by a French press for three consecutive cycles at 10 000 PSI pressure at room temperature. Cell debris was removed by centrifugation with a JA20 rotor at 13 000 rpm for 10 min. RNA was removed after centrifugation at 14 000 g at 10 000 PSI pressure at room temperature. Cell debris was removed by centrifugation with a JA20 rotor at 10 000 rpm for 45 min. For each 100 μL of cell lysate, 200 μL of glacial acetic acid and 33 μL of 1 M MgCl₂ were added. The precipitated RNA was removed after centrifugation at 14,000g on a table-top Eppendorf microcentrifuge. Excess acid was removed by filtration through a 3 kDa Amicon filter several times until the pH reached 6. The pH was checked to be between 5 and 6, which is suitable for strong cation exchange (SCX) chromatography. The protein concentration was determined by a Bradford assay using bovine serum albumin as a standard.

Strong Cation Exchange of Cell Lysate Proteins

Approximately 250 μg of whole cell lysate proteins were loaded onto a 4.6 mm ID SCX chromatography column. The following mobile phases were used: Mobile phase A: 6 M urea + 20 mM acetic acid; Mobile phase B: 6 M urea + 20 mM acetic acid + 500 mM NaCl; pH 5. Proteins were eluted with a 90 min piecewise gradient that was optimized for cell lysate protein separation. Twenty fractions were collected using C4 trapping columns using previously described instrumentation. Each of the 20 fractions was desalted using 5% acetonitrile, 0.1% trifluoroacetic acid (TFA) and then eluted using 90% acetonitrile, 0.1% TFA. The fractions were dried in a vacuum evaporator to remove the acid and acetonitrile. The samples were then resuspended in 25 mM NH₄HCO₃ buffer with 0.1% formic acid, and mass analyzed. The control samples were also treated with 0.1% formic acid before mass analysis.

**LC–MS/MS**

One set of the SCX-separated protein fractions was digested and then analyzed with a Waters or Eksigent liquid chromatograph using a 1 mm C18 reversed phase column with a 120 min gradient. The eluent was electrospayed into a ThermoFisher LTQ Orbitrap. By using its HCD capability, lower mass ion fragments such as b₁ ions could be detected that were not typically observable with the LTQ ion trap alone. Likewise, eight sets of amidinated and unlabeled digested fractions were run on an Eksigent capillary liquid chromatograph and ThermoFisher LTQ ion trap mass spectrometer. In these cases a 100 min gradient was employed for each sample.

**Database Searches**

Acquired mass spectrometer RAW files were converted to mgf files using the default parameters in Proteomictools. Converted mgf files were searched by MSGF+ against the *E. coli* reference proteome consisting of 4306 sequences (Supporting Information) and the often-observed contaminant proteins from the common Repository of Adventitious Proteins (http://www.thegpm.org/rap). In total, the database contained 4306 *E. coli* proteins and 116 contamination proteins as well as the 4422 decoy reversed protein sequences to estimate false discovery rate by target-decoy approaches. For the LTQ ion trap data, the searches were carried out with a 1.5 Da precursor tolerance and semitryptic selectivity. Maximum charge was set at 5 and minimum charge at 1. Carbamidomethyl cysteine was set as a fixed modification, while methionine oxidation, peptide N-terminal amidination (+41.027), and lysine amidination (+41.027) were all variable modifications. For the Orbitrap/LTQ data, the precursor tolerance was set as 10 ppm, the instrument type was “High-res” LTQ, and the fragment method is selected as “HCD”. The other parameters were the same as with the LTQ instrument. It is worth noting that MSGF+ does not accept a “fixed” fragment ion tolerance. Instead, it utilizes mass errors in scoring; for example, a peak with error 0.01 Da contributes higher score than another with error 0.2 Da. MSGF+ also selects an appropriate error model based on the “-inst” parameter. For Orbitrap/LTQ data, we used the default parameter of the instrument. For Orbitrap/LTQ, we set the parameter as “-inst 1”.

**Normalization Method from Spectral Counting Label-Free Quantification**

We first obtained the raw spectral counts for each peptide across all samples. Assuming equal peptide abundances in each sample, we scaled the observed spectral count for each peptide in a sample by the sum of the spectral counts for all peptides detected in that sample. This information was used to interpret the relative detectability of amidinated peptides.

**Classification-Based Postprocessing**

We developed a simple postprocessing procedure to predict the probability that a PSM is a correct identification. The procedure involved training a logistic regression model based on the spectral and peptide features. The positive examples contained the top PSM for every spectrum, whereas the negative examples contained the second best PSM. Both groups were selected from the top 10% of identifications, excluding those from the decoy database, to ensure that the top PSM candidates were correct. The predictor was then applied to top PSMs, and the
target-decoy approach was used to select the score threshold corresponding to the chosen FDR. The details of this workflow are shown in Figure 3A.

The features used to represent peptide-spectrum matches (PSMs) were exclusively related to the annotated fragmentation spectra. They included: (1) precursor error (in ppm units), (2) the charge of the matched spectrum, (3) a binary indicator of a b1 ion match that was set to 1 if the intensity was higher than 5% of the maximum peak intensity, (4) b1 ion match intensity relative to the maximum intensity, (5) a binary indicator of a b2 ion match that was set to 1 if the intensity was higher than 5% of the maximum intensity, (6) b2 ion match intensity, (7) a binary indicator of a yN−1 ion match that was set to 1 if the intensity was higher than 5% of the maximum intensity, (8) yN−1 ion match intensity, (9) b ion matches intensity other than b1 and b2, (10) the fraction of total intensity of the b ion matches other than b1 and b2, (11) the total intensity of the y ion matches other than yN−1 and (12) the fraction of total intensity of the y ion matches other than yN−1. In total, there were 12 numerical features used for predictor development.

To estimate different feature combination effects, we compiled the above-mentioned 12 features into three different groups: b1-including, b1-excluding, and b1-only. The b1-including feature set included all 12 features, the b1-excluding feature set excluded b1, b2, and yN−1 features and the b1-only feature set only kept the b1 feature, specifically only consisted of b1 and yN−1. The b1-including and b1-excluding combinations kept the precursor and charge features. The precursor and charge features were excluded from the b1-only experiments.

Detectability of Amidinated Peptides

We developed another logistic regression classifier to study the signatures of peptides whose detectability is increased after SMTA labeling compared with their unlabeled counterparts. The features used to represent the identified peptides in this model were as follows: (1) a binary vector of length 20 representing the N-terminal amino acid, (2) a binary vector of length 20 representing the C-terminal amino acid, (3) a length-20 vector of amino acid compositions (relative frequencies), (4) molecular weight of the peptide sequence, (5) aromaticity of the peptide sequence,37 (6) hydrophobicity value based on the KD scale,38 (7) hydrophobicity value based on the HW scale,39 (8) hydrophobicity value based on EM scale,40 (9) instability index of the peptide sequence,41 (10) flexibility of peptide sequence,42 (11) isoelectric point of the peptide, (12) fraction of residues in peptides that tend to form helix (amino acids V, I, Y, F, W, L), turn (amino acids N, P, G, S), and sheet (amino acids E, M, A, L) secondary structure, and (13) length of the peptide sequence. In total, we engineered 89 numerical features for the predictor development.

RESULTS

Workflow Overview

Figure 1 graphically displays the experimental protocol. Proteins were extracted from the whole cell lysate of E. coli, then fractionated into 20 traps by ion exchange chromatography in order to reduce protein complexity. Tryptic digestion of the contents of each trap was followed by reversed-phase HPLC–MS/MS analysis. Two types of mass spectrometers were employed to analyze the tryptic peptides. The first, an LTQ-Orbitrap with HCD ion fragmentation, enabled the detection of all fragment ions, including b1 ions. The second, a standalone LTQ ion trap, had a low-mass cutoff that prevented the observation of most b1 ions. The two types of instruments were utilized to explore the overall characteristics of SMTA labeling and its effect on peptide fragmentation.

Improving Peptide Identification by Incorporating b1 Ions

In previous work we demonstrated that peptide amidination can significantly affect the mass distribution of fragment ions generated by collisional activation.27,28 In fact, b1 and, to a lesser degree, yN−1 ions become dominant features in MS/MS spectra of labeled peptides. This is unusual because b1 ions are relatively unstable38,43,44 and are often not observed when unmodified peptide ions are fragmented. Their dominant appearance in the fragmentation of amidinated peptides has been explained by the creation of a cyclic intermediate involving the amidino modification.27 This facilitates peptide backbone cleavage in a manner analogous to the effect that aspartic acid has in inducing charge-remote peptide fragmentation.30 A complication to exploiting this phenomenon is the low-mass cutoff of ion trap instruments that can obviate the observation of b1 ions. Our previous solution to this problem was to employ a TOF mass analyzer that had no such limitation. In the present experiments we utilize an Orbitrap instrument with high-energy collisional dissociation (HCD) ion...
fragmentation to enable the detection of low $b_1$ ion masses. We are thus able to investigate how peptide ion fragment distributions in a large-scale HCD experiment are affected by the amidination modification and whether $b_1$ ions can specifically be exploited to improve peptide identifications.

The numbers of identified spectra in labeled and unlabeled experiments were comparable. In orbitrap labeled experiments, we acquired 1,370 raw spectra. By using target-decoy approaches, based on 1% FDR filtering strategy, 15,755 unique peptides were identified and 1,683 peptides were matched to the decoy database with an MSGF+ E-value cutoff of 2.26 $\times$ 10$^{-8}$ (i.e., 1% FDR). In orbitrap unlabeled experiments, 17,664 raw spectra were recorded and 1,767 unique peptides were identified with 1% FDR with an E-value cutoff of 2.16 $\times$ 10$^{-8}$ (176 PSMs were matched to the decoy database). In terms of unique peptide identifications based on peptide sequence, not on charge states or PTMs, 2,408 unique peptides were identified in SMTA experiments, 2,847 were identified in unlabeled experiments, 1,683 peptides were identified in both experiments, while 725 peptides were identified in the SMTA sample that were not identified in the unlabeled sample (an increase of 25.5%).

**Altered Fragmentation Pattern of Amidinated Peptides.** Of the 15,755 unique peptides identified in SMTA experiments, 15,516 (88.3%) were identified as N-terminally amidinated peptides. We focus on N-terminally amidinated peptides here because the $b_1$ ion is predominantly observed following N-terminal amidination. When annotating spectra of both amidinated and unlabeled samples, a $b_1$ ion assignment was considered to be credible only when the experimentally observed intensity was at least 1% of the maximum intensity in the spectrum. Among the 15,516 unique peptides from the amidinated data, 14,126 (91.0%) unique peptides included a matched $b_1$ ion. This fragment was often intense, accounting on average for 87.9% of the maximum peak intensity in the spectrum. In contrast, of the 17,176 unique peptides from unlabeled peptide experiments, 17,007 (99.0%) were identified as nonamidinated and 169 (1.0%) were identified as N-terminally amidinated or lysine-amidinated. We attributed these 169 amidinated spectra to incorrect matches (Supporting Information).

We also annotated other ions ($b_y$, $y_{N-1}$, $y_{N/2}$) to compare their relative intensities, where $y_{N/2}$ indicates a middle $y$ ion, which dominates the intensity in unmodified peptides. We demonstrate this significant difference in a violin plot (Figure 2). Violin plots are a combination of a box plot and a kernel density plot. Figure 2 shows density of $b_1$, $b_2$, $y_{N-1}$, and $y_{N/2}$ ions in amidinated and unlabeled peptides. Specifically, $b_1$ ions contribute on average 70.1% of the maximum fragment ion intensity in SMTA spectra. Conversely, only 2.3% of the intensity of unmodified spectra are annotated as $b_1$ ions. $b_2$ also shows an obvious difference: 46% of the intensity of SMTA spectra are annotated as $b_2$, compared with 21.5% of the intensity in unlabeled spectra. It is noted here that this significant intensity difference in $b_1$ is the reason to exclude $b_2$ from the feature set $b_1$-excluding. Because of the ease of detecting $b_1$ ions, complementary $y_{N-1}$ ions are also expected, and this is validated in this statistical data. $y_{N-1}$ ions have average 10.8% intensity in SMTA labeled samples, while in unmodified samples they only have 4.5% average intensity. The average intensity of $y_{N/2}$ ions is not significantly affected by amidination. In unlabeled peptides they have an average intensity of 28.5%, while in amidinated peptides their average intensity is 24.3%. In conclusion, our fragmentation data demonstrate that the SMTA did not significantly alter the $y$ series fragment-ion vectors or the $b$ series ion intensities except for $b_1$, $b_2$, and $y_{N-1}$.

**Reproducibility of Amidinated Peptides.** We also investigated the reproducibility of fragmentation patterns in SMTA versus SMTA, unmodified versus unmodified, and SMTA versus unmodified experiments. We annotated the spectra by $b$ series, $y$ series, $b$-$H_2O$ series, $y$-$H_2O$ series, $b$-$NH_3$ series, and $y$-$NH_3$ series. For each unique peptide, we then obtained a set of fixed-length vectors of dimension $6(N - 1)$, where $N$ is the length of the peptide sequence, of annotated ion intensities (a fragment ion was assigned the intensity of 0 no experimental peak could be matched within the $m/z$ tolerance). On the basis of these annotated fragment-ion vectors, we calculated the Pearson correlation coefficients between each pair of spectra identified as the same peptide. The average correlation between all replicated SMTA spectrum-pairs was 0.94 with standard error 0.0016 over 2,131 unique peptides that were identified by at least two spectra, whereas the correlation between replicated unmodified experiments, consistent with our previous observations, was 0.92 with standard error 0.002 over 1,805 unique peptides. Using the set of 1,683 unique peptides that were identified in both SMTA and unmodified experiments, we found that the average correlation coefficient.
between the SMTA and unmodified experiments was only 0.45 with standard error 0.0055 (the reproducibility for a set of unrelated unmodified peptides of the same length was 0.42, whereas the corresponding reproducibility for the SMTA peptides was 0.54). The P value for the difference in means between unmodified versus unmodified and SMTA versus unmodified groups was zero. In summary, the high reproducibility of SMTA spectra and a very low agreement between SMTA and unmodified spectra provide evidence that SMTA labeling results in altered yet reproducible peptide fragmentation patterns.

Peptide-Spectrum Match Postprocessing

On the basis of the tendency of amidinated peptides to form b1 ions, we adopted a supervised approach, with features designed to incorporate b1 ions into a postprocessing procedure (Figure 3A). Prior to classifier development, regular database searching was performed that resulted in a ranked list of peptides for each experimental spectrum. The top PSMs from most confident 10% of identifications were treated as positive examples, while the second best PSMs from the corresponding identifications were selected as negative examples (all decoy hits were excluded). This method was designed to estimate the probability that a PSM reported by a database search engine is correct, for both SMTA and unlabeled experiments, for each feature set (b1-including, b1-excluding, and b1-only; see Methods). The three classifiers, based on logistic regression, were used to better understand the importance of b1 ions for classification.

Prediction Accuracy. To evaluate classification models, we estimated the area under the Receiver Operating Characteristic (ROC) curve of each classifier through 5-fold cross-validation. The area under the ROC curves was estimated as 0.86 for b1-only, 0.96 for b1-including, and 0.94 for b1-excluding for the SMTA experiments and 0.80 for b1-only, 0.95 for b1-including, and 0.94 for b1-excluding for the unlabeled experiments. Details of the cross-validation results are shown in the Supporting Information.

Improving Peptide Identification by Postprocessing.

After scoring each PSM, we set out to apply these scores in the postprocessing procedure, as shown in Figure 3A. Given a particular threshold cutoff τ, we obtained two groups of peptides: one with the prediction score greater than or equal to the cutoff τ, while the other group has scores lower than the cutoff τ. Following this step, we applied a 1% FDR control using target-decoy approach on each of these two groups separately to obtain the final identifications. The underlying rationale for this procedure is that the group with higher prediction probability has a lower chance of having decoy identifications. In Figure 3B, we plotted the MSGF+’s − log (E value) as a function of the prediction score. Forward hits are depicted in black and the decoy hits are in red. Figure 3B shows that the density of decoy hits is very high in the area where the prediction score is low. As the prediction scores gets higher, the density of decoy hits becomes lower.

As shown in Figure 3C, the dependence of the number of identifications on the probability cutoff τ varies for different prediction models. Specifically, by using the b1-including model
and the prediction threshold of \( \tau = 0.4 \), \( \sim 17.2\% \) more PSMs are identified at 1% FDR. The other models, \( b_1 \)-excluding and \( b_1 \)-only, were only able to increase the number of identifications by \( \sim 12.0\% \). We also applied the same strategy on the unlabeled peptide experiments. The results show that \( b_1 \)-only model has no impact on the identification, while the other two (\( b_1 \)-including and \( b_1 \)-excluding) exhibit similar trends, approaching 12% more identified PSMs. As expected, by incorporating the dominant \( b_1 \) and \( b_1 \)-related features in the postprocessing, identifications could be significantly improved in SMTA MS/MS experiments.

While the number of PSMs has been significantly improved by our postprocessing approach, we also compared the unique peptides identified by the original searching (MSGF+ results without postprocessing, which will be referred to as “baseline”) and prediction-based postprocessing. Figure 4 depicts Venn diagrams that indicate the number of unique peptides in each of these four data sets: UnlabelO represents the baseline IDs (unique peptides) in the unlabeled data set, whereas UnlabelP represents the IDs from the unlabeled data set after applying postprocessing. Similarly, SMTAO and SMTAP represent the baseline and the post-processing procedure applied to SMTA experiments, respectively. Figure 4A shows the number of unique peptides supported by at least one spectrum and Figure 4B shows the number of unique peptides supported by at least two spectra. After the postprocessing of the SMTA data set, 227 unique peptides not seen in the unlabeled experiments were identified, 134 of which had at least two PSMs supporting each peptide. The number of identified unique peptides also increased after applying postprocessing to the unlabeled data set, with 258 peptides supported by one spectrum and 174 peptides supported by at least two spectra in the unlabeled data set. Thus postprocessing increased the number of unique peptides in both unlabeled and SMTA experiments, with a slightly higher percent improvement in the SMTA experiments (12.5% in SMTA vs 9.6% in unlabeled), suggesting that incorporating the fragment ion information into the post-processing is beneficial.
Peptide Identification in MS/MS Spectra without b1 Ions

To investigate whether amidination improves peptide identification even when b1 ions are not exploited, peptides were analyzed using an LTQ ion trap mass spectrometer. Whole-cell lysate protein mixtures digested by trypsin were separated into two batches. One batch followed the regular experimental procedure (as described in Methods), while the other was amidinated by SMTA. Both peptide batches were analyzed using LC−MS/MS, with CID-fragmentation in the ion trap. These pairs of experiments were repeated eight times, resulting in 16 different LTQ−LC−MS/MS runs. Peptide identifications obtained in amidinated and unlabeled experiments were compared.

In total, 1,617,477 PSMs were identified from LTQ experiments, from which there were 1,621 decoy hits with an E-value cutoff at 1.1 × 10−5; 196 unique peptides and 1,503 proteins were identified under 1% FDR at the PSM level. Analysis of E. coli whole-cell lysates showed that SMTA labeling enabled the identification of different peptides: 5,342 peptides from nonamidinated experiments, 4,068 peptides in SMTA experiments, and 2,648 in both experiments. Amidination thus enabled the identification of 1420 (26.6%) additional peptides in the cell lysates that were not identified without this modification. Because the LTQ instrument has lower mass resolution and thus gives less confident identifications, we considered peptides supported by at least two PSMs to estimate the increase in confidence by SMTA labeling. 4032 unique peptides were identified (supported by two or more PSMs) in unlabeled experiments, 3296 were identified in SMTA experiments, and 2648 unique peptides were identified in both experiments. 648 additional unique peptides were identified in the amidination experiments, corresponding to a 16.1% improvement. Hence, we conclude that amidination of tryptic peptides could alter both the fragmentation and the detectability of peptides, resulting in the identification of different peptides in MS/MS spectra.

Figure 5 further demonstrates the ability to detect different peptides through SMTA. The number of unique peptides identified in unlabeled experiments increases as more replicated experiments were conducted, but the rate of increase goes down gradually. On the contrary, the slope increases considerably when SMTA data are included. Specifically, after seven unlabeled experiments were combined, the eighth unlabeled experiment was only able to identify 3% more unique peptides, whereas one batch of SMTA experiment can identify 7% additional unique peptides. On the basis of the eight replicated experiments on unmodified peptides, we used a logarithmic function to forecast the incremental number of unique peptides with the increasing number of replicated experiments. The fitted function is expressed as $y = 1606.3 \cdot \ln(x) + 2225.6 \ (R^2 = 0.9934)$ in which $x$ represents the number of replicated experiments and $y$ represents the total number of identified unique peptides. By using this logarithmic function (Figure 5) to extrapolate additional increases in unlabeled experiments, it is evident that SMTA enhances peptide identification, while the replicated unlabeled experiments approach saturation after a small number of experiments.

Quantity Change in Amidinated Peptides

We observed that between 16.1 and 26.6% different peptides were identified in eight replicated LTQ experiments supported by either one or two spectra. This limited increase is reasonable because of the inability to detect b1 ions in the LTQ instrument. In contrast, the Orbitrap instrument benefits from the uniqueness of b1 ions. Nevertheless, even with the LTQ instrument’s inability to detect b1 ions, amidination still enabled more peptides to be detected. Altered ionization efficiency and modulation of the fragment distribution are the two sources of improvement. Moreover, the ionization efficiency alteration contributed not only to the unique peptide identifications but also to the spectral count of detected peptides.

We further examined the peptides that exhibited significant quantitative changes with SMTA compared with unlabeled
experiments. These results are displayed as a heatmap in Figure 6, where the horizontal axis represents 16 (eight unlabeled and eight amidinated) experiments and the vertical axis represents identified unique peptides. The red color represents higher quantity and the blue color represents lower quantity based on spectral counts. We first normalized the label-free spectral counts for each peptide (see Methods). Student’s t test was then used to check whether the quantities were significantly different in SMTA experiments and unlabeled experiments. The P value of 0.05 was chosen as a cutoff to select the peptides, resulting in 212 peptides showing higher quantities in SMTA experiments and 119 peptides having higher quantities in unlabeled experiments (Figure 6). Although both experiments were loaded with equal amounts of proteins, different peptides show different quantitative characteristics. Hierarchical clustering visibly distinguishes the SMTA versus unlabeled experiments through the quantities of identified peptides, indicating that amidination alters the detectability of some peptides. When multiple experiments are run, this process diversifies the set of identified peptides and prevents saturation related to repeated experiments.

We next investigated the characteristics of peptides that were differentially detected in the SMTA and unlabeled experiments. On the basis of these selected peptides, we trained a logistic regression classifier with the features, as described in Methods to predict whether amidination increases or decreases their detected quantity. The area under the ROC curve of this classifier was 76%, which indicates that this is a good classifier (see Supporting Information for the performance of the classifier). We also examined the top ten features with greatest contribution to classification accuracy. As shown in Table 1, the N-terminal residues played a vital role in the classifier, with residues G, E, D, and N having positive contributions and residues V and F having negative contributions. As expected,

![Color Key](image-url)

**Table 1. Regression Coefficients of the Top Ten Features Contributing to the Classifier of Peptides with Differentially Detected Quantities in SMTA and Unlabeled Datasets**

<table>
<thead>
<tr>
<th>feature</th>
<th>regression coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>N-terminal G</td>
<td>1.32</td>
</tr>
<tr>
<td>N-terminal V</td>
<td>-1.08</td>
</tr>
<tr>
<td>N-terminal E</td>
<td>1.08</td>
</tr>
<tr>
<td>N-terminal F</td>
<td>-1.04</td>
</tr>
<tr>
<td>C-terminal R</td>
<td>0.95</td>
</tr>
<tr>
<td>grand average of hydropathy (GRAVY)</td>
<td>0.91</td>
</tr>
<tr>
<td>hydrophobicity of second C-terminus amino acid</td>
<td>-0.84</td>
</tr>
<tr>
<td>N-terminal D</td>
<td>0.82</td>
</tr>
<tr>
<td>hydrophobicity of first N-terminus amino acid</td>
<td>-0.72</td>
</tr>
<tr>
<td>N-terminal N</td>
<td>0.71</td>
</tr>
</tbody>
</table>

Figure 6. Heatmap of quantities of identified unique peptides in unlabeled and SMTA across 16 experiments. The x axis is the 16 experiments (8 unlabeled and 8 SMTA) and the y axis represents the identified unique peptides. The color reflects the normalized spectral counts quantity: blue indicates lower quantity and red indicates higher quantity.
hydrophobicity-related characteristics of the peptides were important features for the classifier. Because the amidinated peptides are more basic than unamidinated peptides, this labeling is expected to alter detectability of peptides, depending on their amino acid composition and overall hydrophobicity.

### SUMMARY AND CONCLUSIONS

Our prior experience led us to investigate the impact of SMTA labeling of peptides on protein identification in high-throughput shotgun proteomics. By introducing an amidino group into the N-termini of peptides, their fragmentation patterns and detectability have been significantly altered. We observed that $b_1$ ions are primarily and significantly enhanced, which allows accurate recognition of peptide N-terminal amino acids. Accordingly, the complementary $y_{N-1}$ ions are enhanced to a lesser degree, while $b_2$ fragment ions are diminished. Even though $b_1$ and $y_{N-1}$ ions are most significantly intensified in the MS/MS spectra of amidinated peptides, the intensities of $y_{N/2}$ fragment ions are hardly affected. This is a notable feature that can be exploited in targeted proteomics such as selective reaction monitoring (SRM) for protein quantification.

On the basis of the 16 replicated LTQ experiments in which $b_1$ ions were not detected, we found that SMTA modification not only enables the identification of additional peptides but also significantly changes the fragmentation patterns of some peptides. We reasoned that the detectability of these peptides is altered by the increased basicity of their N-termini. Understanding the characteristics of the peptides with altered detectability could help to identify and quantify some peptides that are unlikely to be observed in experiments with unamidinated samples.

Finally, in this work, we also showed that integrating prior knowledge of amidinated peptide fragmentation into the computational postprocessing can substantially improve peptide identification. Overall, we estimated that the postprocessing procedure implemented in this study identifies 12.5% additional peptides in SMTA experiments that are unlikely to be identified otherwise.

### ASSOCIATED CONTENT

* Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.6b00468.

Supplementary Methods. Supplementary Figure S1: Annotated spectra for example amidinated peptide ID. Supplementary Figure S2: ROC for $b_1$-including in SMTA data. Supplementary Figure S3: ROC for $b_1$-excluding in SMTA data. Supplementary Figure S4: ROC for $b_1$-only in SMTA data. Supplementary Figure S5: ROC for $b_1$-including in unlabeled data. Supplementary Figure S6: ROC for $b_1$-excluding in unlabeled data. Supplementary Figure S7: ROC for $b_1$-only in unlabeled data. Supplementary Figure S8: ROC for detectability change. (PDF)

### AUTHOR INFORMATION

Corresponding Authors

*P.R.: E-mail: predrag@indiana.edu.

*J.P.R.: E-mail: reilly@indiana.edu.

### ACKNOWLEDGMENTS

This work was supported by the NIH grant R01 GM103725. S.A.M. was supported by an NIH grant to the Protein Translation Network (U54 GM105816).

### REFERENCES


Notes

The authors declare no competing financial interest.
(32) Li, Y. F.; Arnold, R. J.; Tang, H.; Radivojac, P. The importance of peptide detectability for protein identification, quantification, and experiment design in MS/MS proteomics. *J. Proteome Res.* 2010, 9, 6288−6297.
(34) Liu, X.; Reilly, J. P. Correlating the chemical modification of Escherichia coli ribosomal proteins with crystal structure data. *J. Proteome Res.* 2009, 8, 4466−4478.