analytical chemistry

Article

pubs.acs.org/ac

pTop 1.0: A High-Accuracy and High-Efficiency Search Engine for Intact Protein Identification

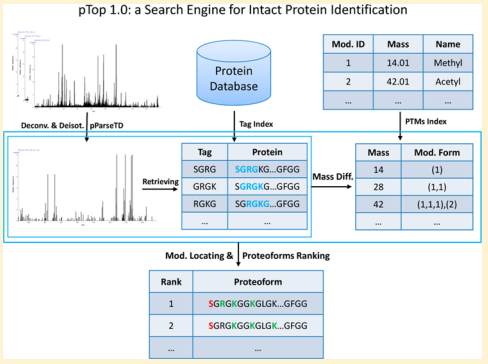
Rui-Xiang Sun,*[,†] Lan Luo,[†,‡] Long Wu,[†,‡] Rui-Min Wang,[†,‡] Wen-Feng Zeng,[†,‡] Hao Chi,[†] Chao Liu,[†] and Si-Min He[†]

[†]Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China

[‡]University of Chinese Academy of Sciences, Beijing 100049, China

Supporting Information

ABSTRACT: There has been tremendous progress in top-down proteomics (TDP) in the past 5 years, particularly in intact protein separation and high-resolution mass spectrometry. However, bioinformatics to deal with large-scale mass spectra has lagged behind, in both algorithmic research and software development. In this study, we developed pTop 1.0, a novel software tool to significantly improve the accuracy and efficiency of mass spectral data analysis in TDP. The precursor mass offers crucial clues to infer the potential post-translational modifications co-occurring on the protein, the reliability of which relies heavily on its mass accuracy. Concentrating on detecting the precursors more accurately, a machine-learning model incorporating a variety of spectral features was trained online in pTop via a support vector machine (SVM). pTop employs the sequence tags extracted from the MS/MS spectra and a dynamic programming algorithm to accelerate the search speed, especially for those spectra with multiple post-translational modifications. We tested pTop on three publicly available data sets and compared it with ProSight and MS-Align+ in terms of its recall, precision, running time, and so on. The results showed that pTop can, in general, outperform ProSight and MS-Align+. pTop recalled 22% more correct precursors, although it exported 30% fewer precursors than Xtract (in ProSight) from a human histone data set. The running speed of pTop was about 1 to 2 orders of magnitude faster than that of MS-Align+. This algorithmic advancement in pTop, including both accuracy and speed, will inspire the development of other similar software to analyze the mass spectra from the entire proteins.



Top-down proteomics (TDP), focusing primarily on large-scale intact proteins, has been gaining in popularity over the past five years. TDP encompasses front-end fractionation, mass spectrometry, and back-end bioinformatics.[1−4] During the pioneering days, only the purified protein or very simple samples could be resolved.[2] To date, however, more than 1000 proteins in complex samples can be identified in just a single study. Indeed, top-down protein analysis has been achieved on the proteomic level.[1,5−7] Under today's technology-driven circumstances, a great variety of publications employ TDP, either in fundamental biological research[8,9] or in clinical and translational applications.[10,11] For instance, Ge's lab recently discovered three crucial cardiac proteins in acutely infarcted swine myocardium and then comprehensively sequenced these proteins and pinpointed their phosphorylation sites by top-down mass spectrometry.[12]The great demand for TDP in biomedical applications has pushed TDP toward investigating more complex proteins.

Protein complexity lies not only in the protein itself, but also in its upstream gene and transcript. A new term, "proteoform," frequently used in TDP, is defined as "all of the different molecular forms in which the protein product of a single gene can be found, including changes due to genetic variations, alternatively spliced RNA transcripts and post translational modifications".[13] TDP is becoming a powerful tool for identifying the large-scale proteoforms, whereas the widely available bottom-up proteomics (BUP) concentrates on digested peptide mixtures.[14] The protein inference problem[15−17] in BUP can cause ambiguities when determining which proteins actually exist in the sample from the peptide pool. Such ambiguities may also be carried forward into protein quantification. Nevertheless, TDP bypasses the digestion process and protein inference in BUP. It aims to identify and characterize the proteins from a "bird's eye" view (i.e., all proteoforms globally).

So far, TDP has evolved from a promising technique adopted in only a few research laboratories to a practical tool applied in a great many biological investigations. This can be attributed mainly to the remarkable advancements made in both fractionation and mass spectrometry. An effective separation procedure is crucial to achieve large-scale intact protein identification. In 2011, Kelleher's lab developed a new four-

Received: October 19, 2015
Accepted: February 4, 2016
Published: February 4, 2016

DOI: 10.1021/acs.analchem.5b03963
Anal. Chem. 2016, 88, 3082−3090

dimensional, liquid-phase separation system, which increased the separation power 20-fold over any previous work. Based on this system, 1043 gene products from human cells were then identified, each with about three proteoforms discovered, on average.[1] To date, the largest top-down study was reported to have identified 1220 proteins from a human cell line H1299, which carried out multiple separation strategies, also from Kelleher's lab.[5] It was even regarded as a milestone in TDP that the barrier of the 1000-protein scale had been broken—implying that even larger-scale identification would not be far behind.[7] In addition to separation, intact protein identification depends heavily on mass spectrometry (MS), particularly the high-performance MS, such as high-resolution, high-accuracy Fourier Transform (FT)-based instruments. In recent years, both the resolution and the speed have been improved dramatically, leading to higher-quality MS and MS/MS data than those produced from the preceding versions. There are also different, but complementary, fragmentation methods available to sequence the whole proteins, such as HCD,[18] ETD,[19] ECD,[20] the new, promising UVPD,[21] and so on. All of this technical progress has laid the foundation for proteoform characterizations.

The large-scale and complicated protein MS data pose a highly computational challenge to downstream bioinformatics, which has played an increasingly crucial role in TDP. Unlike BUP, with more than a dozen software tools worldwide, only a few are currently available to deal directly with top-down data. Among them are ProSight,[22−26] the first top-down software, and MS-Align+,[27] another popular alternative. Algorithmic study and software development lag far behind when compared with the dedications to separation and MS. A virtual issue collected the TDP research papers that were published in the Journal of Proteome Research and Analytical Chemistry before June 2013. There were 29 papers on separation and 28 papers on MS, but only six were primary research papers on information for protein identification.[28] In most cases, the manual interpretation was inevitable due to the insufficient tools available or the unreliable output from the software used.

Compared with the peptide spectra, the spectra of intact proteins are much more complicated in terms of the isotopic profile, charge state, number of fragments, and so on. Although it seems daunting to confront such complex spectra to be processed, such spectra abound with valuable information to be discovered. For instance, there are more isotopic peaks as the protein molecular weight becomes larger. Both the charge state and accurate mass can be inferred from its corresponding isotopic profiles if the resolution is high enough to resolve them. Thus, it is important to effectively utilize this sort of information when designing a novel software tool. Moreover, from the viewpoint of computational efficiency, owing to the huge search space when considering the combinatorial post-translational modification (PTM) patterns on the protein of interest, the analysis of large-scale and complicated mass spectra still remains low-efficiency. Current software has much room to improve not only because TDP itself needs a much larger search space than that of BUP but also because mass spectra of intact proteins require more preprocessing computation. Thus, how to improve preprocessing and searching efficiencies is another factor to be considered. We should balance the trade-off between sensitivity and running time. Spectra preprocessing could reduce the complexity and improve the sensitivity of precursor detection. Enlarged search space could also increase the number of identified spectra, but this requires the high-level algorithms to achieve the high-efficiency, especially when searching the proteins harboring multiple PTMs, such as the histones.

Generally, the first step to process top-down MS data is to convert each isotopic cluster to its equivalent monoisotopic mass with its absolute or relative intensity (i.e., spectral deconvolution). In 2000, Horn et al. proposed an algorithm, THRASH,[29] to handle the deconvolution of the high-resolution mass spectra of intact proteins. THRASH uses a subtractive peak-finding method to iteratively pick the isotopic clusters and converts them to the corresponding monoisotopic masses. The formula to calculate the signal-to-noise ratio is now adopted in many systems. As the first algorithm to process the mass spectra of intact proteins, it has been extended in Decon2LS,[30] DeconMSn,[31] Xtract,[25] and MASH Suite.[32] However, its processing speed did not keep a pace with mass spectrometry because the speed of mass spectra acquisition has been improved significantly in the past decade. In 2010, Liu et al. released another preprocessing software, MS-Deconv,[33] which put forward a combinatorial algorithm for spectral deconvolution in order to select a highest-scoring subset of isotopic clusters rather than iteratively selecting the highest-scoring isotopic clusters using a greedy algorithm. However, it does not support the export of all the precursors if they are coeluted.

After such spectral preprocessing, the second step is usually to identify the proteins using the database search strategy. Because the sequence of an intact protein is, on average, much longer than that of a peptide and the potential modifications may be much more complex than those on a peptide, enumerating each possible variable modification may cause a combinatorial explosion. For example, if all possible modifications on human histone H3.1 protein in the annotated database are listed, the number of possible proteoforms will reach 40 trillion in theory.[17] As the first search engine designed for identification of intact proteins, ProSight[22−25] adopts an offline database strategy—'shotgun annotation'—to generate a virtual database including known proteins, splice variants, and modifications, and then it searches the virtual database online. ProSight was the only software available to identify proteoforms from tandem mass spectra of intact proteins until 2008.[34] However, it has one shortcoming that there are a large number of proteoforms in the virtual database that will not be actually matched. The size of the virtual database will become huge for complex species. It is essentially difficult to scale, as Pevzner pointed out.[35] MS-Align+,[27] another popular tool in TDP, was developed on the basis of a spectral alignment algorithm from MS-TopDown.[36] Because the mass deviation during the alignment between the mass lists of a spectrum and a protein can be arbitrary, MS-Align+ can search the unexpected modifications using a dynamic programming algorithm. However, as the number of potential modifications increases, the time and space required are still very large. MS-Align+ has introduced several tricks to accelerate identification.[27,37] In addition to these two popular search algorithms, there are a few other algorithms for TDP, such as PIITA[38] and ProteinGoggle.[39]

In this work, we developed a novel software tool, pTop 1.0, to address current low-efficiency computational issue in TDP. In pParseTD, a preprocessing module in pTop, to detect the precursors more accurately, a model with multiple characteristics of the precursor ions was incorporated by online training in a machine-learning approach. pParseTD could not only recall more accurate precursors but also support the export of

DOI: 10.1021/acs.analchem.5b03963
Anal. Chem. 2016, 88, 3082−3090

the multiple precursors in the coeluted spectrum. In the database search procedure, we adopted a sequence-tag based protein identification strategy which was similar as that in ProSight. The difference was that we introduced the indexes of both sequence tags and combinatorial modifications in order to accelerate the searching process. Through two-step retrieval in the indexes, the search space of candidate proteoforms could be reduced substantially. Finally, we designed a delicate scoring approach based on dynamic programming to locate the modifications and simultaneously match the spectrum with all candidate proteoforms derived from a certain combination of modifications. pTop was tested on three public data sets. The preprocessing experimental results on the human histone data set showed that pParseTD could recall 22% more correct precursors while exporting 30% fewer precursors than that of Xtract, on average. Meanwhile, pTop could achieve a speed of up to about 1 to 2 orders of magnitude faster than that of MS-Align+ on the complex data sets tested. pTop can be downloaded for free at http://pfind.ict.ac.cn/software/pTop/index.html.

MATERIALS AND METHODS

General Workflow of pTop. pTop, as a novel database search engine to identify intact proteins, aimed at significantly improving search efficiency while retaining higher accuracy. Its general workflow was designed as shown schematically in Figure 1, which consists of four main procedures according to the data streaming direction:

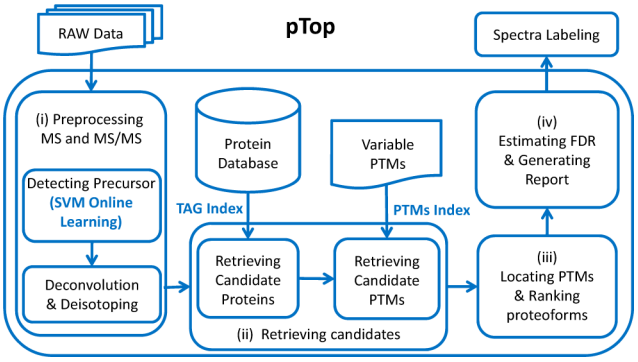


Figure 1. Schematic workflow of pTop.

(i) preprocessing raw mass spectral data;
(ii) retrieving candidate proteins and combinatorial modifications;
(iii) locating the modifications and ranking the candidate proteoforms;
(iv) estimating the false discovery rate (FDR) and then generating the result report.

To preprocess both MS and MS/MS raw data, we developed pParseTD—an essential preprocessing module in pTop—to first detect potential precursors within the isolation window in MS, and then to convert each detected isotopic cluster in MS/MS into its equivalent monoisotopic peak with a single charge. This conversion will greatly simplify the subsequent matching process between an MS/MS spectrum and its candidate protein.

For each MS/MS spectrum processed by pParseTD, in order to obtain its candidate proteins from the protein database, the sequence tags were first extracted therein and then searched against the index of sequence tags of the protein database. After that, the candidate combinatorial modifications were generated using the mass difference between the precursor and the candidate protein. Through these two steps, the search space of the proteoforms decreased tremendously, and the identification was greatly accelerated.

The next procedure was to determine the possible locations of such candidate modifications in a coordinated fashion. Here, we designed a delicate scoring function to match the proteoform with the input MS/MS spectrum, which promotes its scoring efficiency by the simultaneous matching for different candidate combinations of the modification sites.

Finally, a target-decoy approach was utilized to estimate FDR and determine the scoring threshold. This approach is widely used in BUP and was also recently adopted in TDP.[27,38] Each MS/MS spectrum preprocessed by pParseTD was searched against the cascaded target and reversed protein database. Then all protein-spectrum matches (PrSMs) filtered by an FDR cutoff of 1% were kept into the result report.

Detection and Deconvolution of Isotopic Clusters. We used pParseTD, a nontrivial updated version of pParse 2.0,[40] which was originally developed to detect the peptide precursors, to detect intact protein precursors in MS. Moreover, pParseTD provided a similar approach for MS/MS deconvolution to convert each original MS/MS spectrum into a list of monoisotopic peaks with a single charge.

As is known, most algorithms for detecting the isotopic clusters were based on the similarity or dot product between the experimental isotopic distribution and the averagine theoretical isotopic distribution.[29,33,41] However, when the intensity of a peak is not high enough, its isotopic cluster may be distorted, which will weaken their similarity. Furthermore, when a few isotopic clusters overlap, the similarity-based comparison will become ineffective or insufficient. Therefore, we used, in pParseTD, a variety of features to characterize and evaluate the candidate isotopic clusters. Those features were obtained mainly from the experimental isotopic clusters, liquid chromatography (LC) profiles, and the additional information from the same protein ion with different charge states (more details on all 11 features in Table S-1 and Figure S-1).

In order to integrate those features to evaluate the isotopic clusters, a machine-learning approach was introduced into pParseTD. A classification model was trained online, based on the support vector machine (SVM) method with a radial basis kernel function. Using this classification model, pParseTD can automatically incorporate a variety of features to detect the isotopic clusters and determine their charge states. This model can adapt to the different input raw data.

In the training and predicting sections, several critical problems must be addressed. First, all candidate precursors were scored using the values of above features, and only the top 10 precursors were retained. Then, the candidate in the first place (Top 1) was selected as a positive sample, and the 10th, or the last one if there are fewer than 10 candidates, as a negative sample. Second, when using an SVM to train the model online, the leave-one-out cross-validation method was employed to reduce the risk of overfitting. Third, after obtaining the model, each candidate isotopic cluster was rescored by the same model, and only those with a predicted score larger than zero were exported. Finally, an offline model in pParseTD would be used if the number of training samples was not sufficient.

DOI: 10.1021/acs.analchem.5b03963
Anal. Chem. 2016, 88, 3082−3090

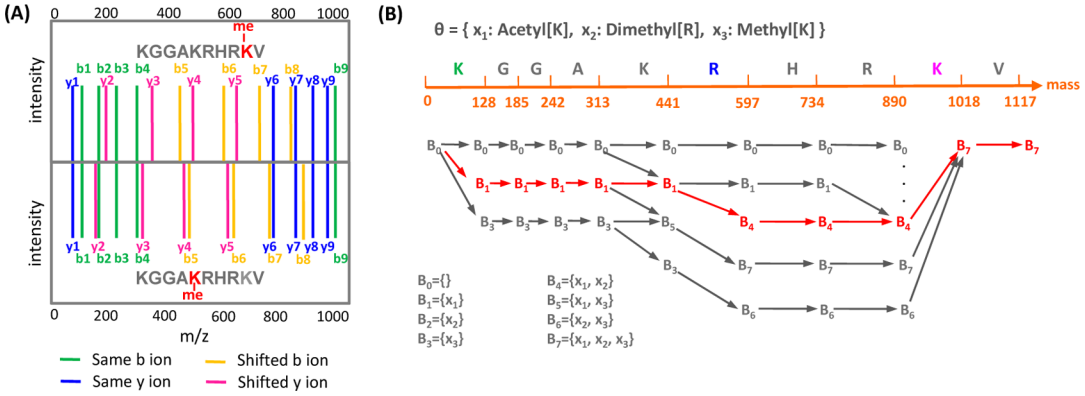


Figure 2. (A) Theoretical spectra of two proteoforms from the protein sequence of KGGAKRHRKV. The top panel shows the proteoform with one methylation on the ninth lysine, and the bottom panel shows the proteoform with one methylation on the fifth lysine. From the two spectra, we can see that only four b ions (golden) and four y ions (rosy) have shifted, while all others are shared. (B) An illustration for locating the modifications by converting the best PrSM to a maximum weighted path in the graph. Suppose that the proteoform of the spectrum is K(ac)GGAKR(me2)-HRK(me)V, which indicates the acetylation on the first amino acid K, the dimethylation on the sixth amino acid R, and the methylation on the ninth amino acid K. Each red vertex in the graph stands for a pair of matched fragment ions. In this case, the maximum weighted path (red) is definite, and from the path, we can find the best matched proteoform of the spectrum.

Retrieval of the Candidate Proteins and Their Modifications. In fact, most candidate proteins could not be matched well with the spectrum if they were chosen according to a mere precursor mass in MS. It is the MS/MS in which the fragment peaks provide much more important information related to the protein identity that can be used to further reduce the number of candidate proteins. Therefore, using the peaks in MS/MS, we first extracted the sequence tags (Figure S-2) in order to retrieve the top 100 candidate proteins. Then using the mass differences between the precursor and the candidate proteins, we generated the candidate combinatorial modifications according to a list of user-defined variable PTMs. Those mass differences imply very important information from which to infer the most possible combinatorial modifications or any other variations. They can limit the number of all combinations, and only those that satisfy the mass constraints are required for later scoring. This kind of beneficial mass-related information employed in pTop will substantially improve search efficiency, particularly for those proteins with multiple modifications, when compared with the "shotgun annotation" in ProSight. To further speed the retrieval of the candidate protein sequences and their modifications, we built indexes of sequence tags and combinations of user-specified variable modifications before searching.

Location of Modifications and Ranking of Proteoforms. As described above, given the candidate protein and its possible combinatorial modifications, we could obtain a few different protein variants—i.e., proteoforms. The differences among them are only the locations of each modification—that is, their theoretical spectra are highly similar. For example, Figure 2A shows that a toy protein sequence is KGGAKRHRKV, and one variable modification is methylation on lysine(K). We could get three proteoforms, which vary in the location of the methylation. Consider two proteoforms with methylation on the fifth lysine or the ninth lysine and their theoretical fragment ions (e.g., b and y ions). We can generate the theoretical spectra of these two proteoforms, as shown in Figure 2A. The y1, y6, y7, y8, y9 ions (blue) and b1, b2, b3, b4, b9 ions (green) are shared by these two proteoforms. Other ions, such as b5, b6, b7, b8 and y2, y3, y4, y5 have shifted due to the mass from methylation. The shared fragment ions from different proteoforms could be calculated and matched only once. Accordingly, we could scan the protein sequence only once and count the matched fragment ions of all candidate proteoforms generated from the given protein and its combinatorial modifications.

The procedure to locate the modifications on the candidate protein is stated as follows. First, we generate the mass list M_P of a candidate protein. A protein P = $r_1r_2...r_m$ is a sequence of amino acids with the length m. The mass list of P is a series of theoretical masses of the fragment ions represented as M_p = $\{b_1^+, b_2^+,...,b_{m-1}^+,y_1^+,y_2^+,...,y_{m-1}^+\}$ for CID or HCD spectrum matching (c and z ions for ECD or ETD). Let S = $\{x_1,x_2,...,x_k\}$ be the set of the variable modifications that user defined, and θ = $\{x_i,x_j,...,x_t\}$ be a set of valid combinatorial modifications, where x_i represents the i^{th} variable modification in S. The elements in θ are repeatable and disordered. For example, θ = $\{x_1,x_1,x_3,x_5\}$ is possible. Given the mass list M_S of an input MS/MS spectrum, the mass list M_P of a candidate protein, and the combinatorial modifications θ = $\{x_i,x_j,...,x_t\}$, a directed acyclic graph (DAG) G is constructed as shown in an example of Figure 2B. The vertex in the graph is defined as (i, B_j), where i represents the location of the amino acid in the protein, B_j denotes the set of modifications on the subsequence from the first amino acid to the i^{th} one, and $B_j \subseteq \theta$. For convenience, we add a vertex $(0, B_0)$ as the source, where B_0 is an empty set. The vertex (m,θ) is always the sink. For each amino acid in the protein sequence, there are several corresponding vertices in the graph which are in the same layer. For each vertex in the graph G, the corresponding weight $w(i,B_j)$ is initialized to zero and added by one, if there is a mass in the list M_S which could match the mass of $\{b_i^+ + m(B_j), y_{m-i}^+ + m(\theta) - m(B_j)\}$ within a given tolerance, where $m(\theta)$ denotes the total mass of the modifications in θ and $m(B_j)$ denotes the total mass of the modifications in B_j. For two adjacent-layer vertices (i, B) and $(i + 1, B')$, they could be connected by a directed edge, if one of the following two conditions is satisfied. (i) $B = B'$; (ii) $B' = B\cup\{x\}$, $x \in \theta$ and the modification x could occur on the $(i + 1)^{th}$ amino acid. The weight of a path in the graph G is defined as the sum of the weights of the traversed vertices. A path is valid if it is connected from the source to the sink, that is to say, every modification in θ occurs on protein P. Since each valid

DOI: 10.1021/acs.analchem.5b03963
Anal. Chem. 2016, 88, 3082−3090

Table 1. Comparisons of the Recalled Precursor Rate (RPR) and the Number of Exported Precursors on the MSD Precursor Benchmark

			Xtract		pParseTD	
data set	#MS/MS[a]	benchmark[b]	#precursor (avg.)[c]	recalled precursor rate	#precursor (avg.)	recalled precursor rate
H2A	7270	2013	30000 (4.1)	37.9%	15708 (2.2)	79.8%
H2B	5863	2879	20155 (3.4)	70.4%	16378 (2.8)	96.1%
H3	6924	1470	38026 (5.5)	70.5%	29244 (4.2)	80.4%
H4	2698	958	6543 (2.4)	88.5%	5000 (1.7)	90.3%
All	22755	7320	94724 (4.2)	66.2%	66330 (2.9)	88.1%

[a]Total number of tandem mass spectra acquired in each raw file in human histone data set. [b]Number of precursors in the MSD precursor benchmark. [c]Total number of precursors exported from the software (#precursor) and the average number of precursors for each tandem mass spectrum (avg.), which is equal to the total number of exported precursors divided by the number of MS/MS acquired (the second column).

path in the graph G represents a proteoform, the best weighted path is considered as the best matching proteoform. The problem of modifications location and proteoforms ranking is now reduced to an optimal path-finding problem. pTop uses the pDAG[42] algorithm to find the k-best paths.

RESULTS

Data Sets and Benchmarks. The performance of pTop 1.0 was tested on three public data sets[21,43,44] (details listed in Table S-2). The main results in this manuscript were based on the human histone data set;[43] the results from the other two data sets were summarized in the Supporting Information.

To evaluate the accuracy of the precursors detected by the preprocessing module, pParseTD, we constructed a precursor benchmark on the human histone data set.[43] The histone MS data contain many high-quality spectra with multiple modifications which are appropriate to test pTop's efficiency under combinatorial modifications. In order to compare the software tools to export multiple precursors for the coeluted spectra, we used MS-Deconv to annotate the precursors in raw files of the human histone data set. To ensure the reliability of precursors exported by MS-Deconv, we used MS-Align+ to identify all MS/MS spectra. Ultimately, 7320 reliably identified precursors were included in the MSD (MS-Deconv) precursor benchmark. To evaluate the performance of deconvolution and deisotoping, we constructed a peak benchmark by peak annotation in MS/MS based on the human histone data set. To generate the peak benchmark in MS/MS, the reliable proteoforms were identified by both ProSight and pTop (the numbers of consistent identifications were shown in Figure S-3). Then, all theoretical fragment ions with different charge states were enumerated, and their isotopic distributions were calculated by an "emass" algorithm.[45] The next step was to align the theoretical isotopic clusters with the experimental ones and then to calculate their similarities. Finally, the isotopic cluster with a similarity score greater than 0.9 was labeled with the charge state of the theoretical one. There were, in total, 69 913 isotopic clusters, which were selected to generate the monoisotopic peaks with a single charge, from 2769 MS/MS spectra. These monoisotopic peaks constituted the peak benchmark.

We first tested the performance of the preprocessing module, pParseTD, on the above two benchmarks and compared it with that of Xtract (in ProSight) and MS-Deconv. The recall rates of the precursors and MS/MS peaks were compared among these three software tools. Then, we compared the performance of pTop with ProSight PC 3.0 and MS-Align+ on the human histone data set in terms of accuracy and running time. Because there are some inherent parameters in each software tool, we chose the parameters that could balance the accuracy and the speed. ProSight searched against the annotated database under the absolute mass mode. MS-Align+, as the open search software, searched against the target and decoy protein databases with the limit of the maximum number of unexpected modifications on each protein. pTop searched against the target and decoy protein databases with a list of user-specified variable modifications and the maximum number of modifications on each protein. The FDR used to filter the reliable PrSMs was set as 1% for MS-Align+ and pTop. For ProSight, the PrSMs with an *E*-value less than 0.0001 (default) were reported as the reliable identifications.

Evaluation of the Accuracy on Precursor Detection. We defined a precursor as recalled if the absolute value of the mass deviation between the benchmark precursor and the precursor given by the preprocessing software was less than 3.1 Da. The recalled precursor rate (RPR) is the number of recalled precursors divided by the total precursor number in MSD benchmark data set. The precision of pParseTD was compared with that of Xtract in terms of the number of exported precursors (i.e., the average number of precursors for each spectrum). We chose the "Top Down (MS2)" option in Xtract to process the raw data. The default settings for this option were listed in Table S-3. The result is shown in Table 1. pParseTD recalled about 22% more precursors while exporting 30% fewer precursors than those of Xtract. That is, the RPR of pParseTD is 22% higher than that of Xtract, although it exported 30% fewer precursors. Similar comparisons were also conducted on the other two data sets. The detailed result is shown in Table S-4. For Autopilot data, the RPR of pParseTD is about 40% higher than that of Xtract. For *E. coli* data, both Xtract and pParseTD can achieve to an RPR of 97% ∼ 98%. This performance difference between these two software tools is related to the sample complexity. We will later analyze it in the Discussion section.

To detect the precursors in MS more accurately, pParseTD not only took the similarity between the theoretical and experimental isotopic clusters into consideration, but also used the similarity of LC profiles and other information about the same protein precursors with different charge states. In addition, a machine-learning method was used in pParseTD to handle the coeluted precursors. The tests revealed that pParseTD achieved higher accuracy than Xtract on the above MSD benchmark.

Evaluation of the Accuracy on Monoisotopic Peaks in MS/MS. To evaluate the accuracy of the monoisotopic masses of the fragment ions in MS/MS, a peak benchmark on the human histone data set was constructed, as described previously. Here, we defined a peak as recalled if the absolute

DOI: 10.1021/acs.analchem.5b03963
Anal. Chem. 2016, 88, 3082−3090

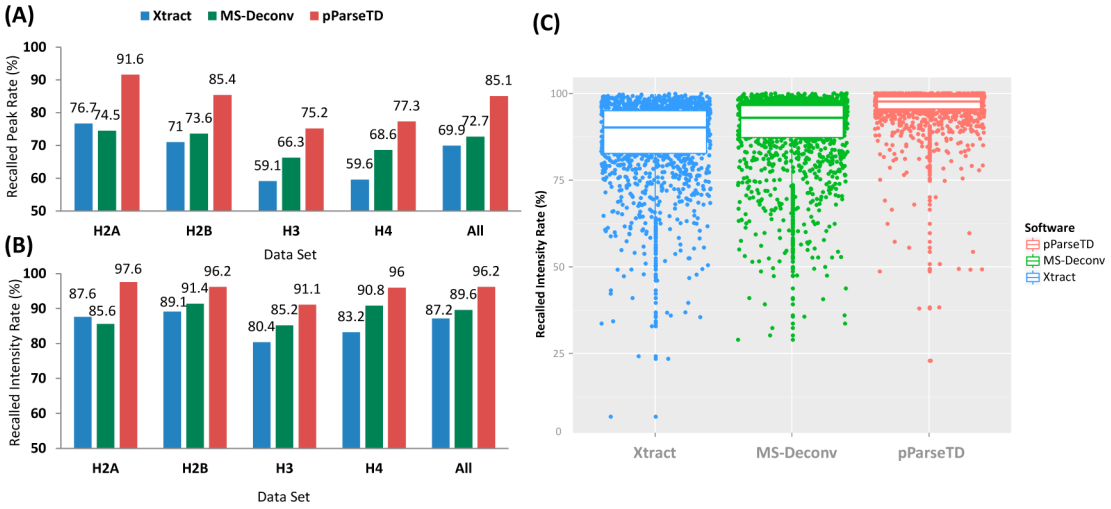


Figure 3. Recalled peak rates (RKR) (A) and recalled intensity rates (RIR) (B) by Xtract, MS-Deconv, and pParseTD in the peak benchmark on human histone data set. "H2A", "H2B", "H3", and 'H4" stand for the four different raw files, and "All" represents the entire data set. (C) Boxplots of RIRs by Xtract, MS-Deconv and pParseTD on 1528 MS/MS spectra. The medians of RIRs by Xtract, MS-Deconv and pParseTD are 90.1%, 93.0% and 97.7%, respectively.

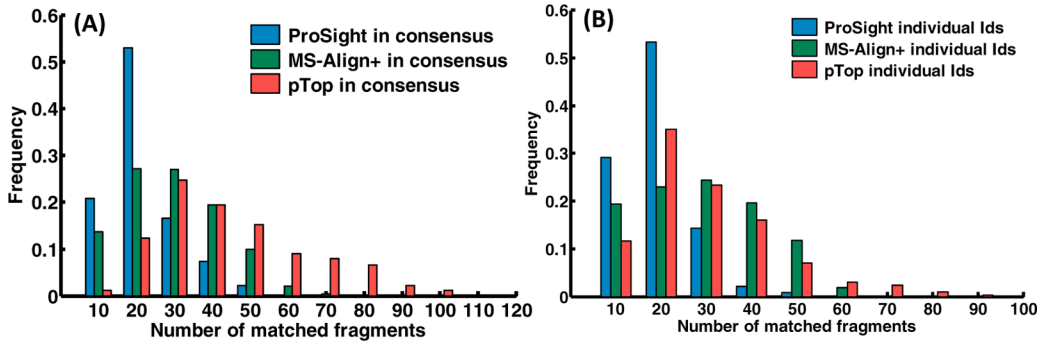


Figure 4. Distributions of the matched fragment ions. (A) The distributions of the number of matched fragment ions in the consensus identifications of ProSight, MS-Align+, and pTop. (B) The distributions of the matched fragment ions in the individual identifications of ProSight, MS-Align+, and pTop.

value of the mass deviation between the benchmark peak and the peak given by the software was less than 20 ppm. The recalled peak rate (RKR) was defined as the number of recalled peaks divided by the total peak number in the peak benchmark data set. The recalled intensity rate (RIR) was defined as the sum of the intensity of the recalled peaks divided by the total intensity of the annotated peaks for each MS/MS spectrum in the peak benchmark data set. The performances of deconvolution and deisotoping of Xtract, MS-Deconv, and pParseTD were analyzed on the peak benchmark. Figure 3A,B show their recalled peak rates and recalled intensity rates, respectively. The experimental results show that pParseTD achieved 85.1% RKR and 96.2% RIR, on average, in this peak benchmark. Moreover, the RKR of pParseTD was about 15.2% and12.4% higher than that of Xtract and MS-Deconv, respectively, and 9.0% and 6.6% higher for RIR. We also carried out these similar comparisons on the other two data sets. The detailed results were summarized in Table S-5. The test experiments show that pParseTD achieved a higher accuracy for more complex samples by integrating a variety of features to recognize the isotopic clusters. The single feature of the similarity between the theoretical and experimental isotopic clusters in Xtract and MS-Deconv might not be high enough when the intensity of the isotopic cluster is low in complex samples. We will also discuss the performance difference on RKR and RIR for different data sets in the Discussion section.

Furthermore, for each tandem mass spectrum in the peak benchmark, Xtract, MS-Deconv, and pParseTD output a value of RIR, respectively. The distributions of these RIRs are illustrated in Figure 3C. The number of spectra with an RIR above 90% occupies 93% in pParseTD, 65% in MS-Deconv, and 51% in Xtract. A higher RIR in pParseTD implies that there are more high-intensity peaks recalled, which reflects a higher confidence level.

Performance Comparison among ProSight, MS-Align+, and pTop on the Human Histone Data Set. The human histone data set was searched by ProSight, MS-Align+, and pTop, respectively. More detailed parameters for the database search were listed in Table S-6. In order to compare ProSight and pTop under the similar condition, here, we only used the single absolute mass search mode in ProSight. For MS-Align+, the shift number was set as the default 2. The numbers of identified spectra are 6144, 9191, and 6240 by ProSight, MS-Align+, and pTop, respectively. Among them, 5847 spectra within the precursor mass deviation in [−2.2, 482.2]Da are to be further analyzed (more details shown in Figure S-4). There are 1821 consensus identifications among the three identification results, which are less than one-third of

DOI: 10.1021/acs.analchem.5b03963
Anal. Chem. 2016, 88, 3082−3090

the total identifications by each software tool. The comparison of identified spectra is nontrivial, so, here, the number of matched fragment ions is analyzed to assess the relative reliability of the identifications.

The numbers of matched fragment ions in the consensus and individual parts were selected for further statistical analysis. Figure 4A shows the distributions of the numbers of matched fragment ions for ProSight, MS-Align+, and pTop in the consensus identifications. pTop matches more fragments than MS-Align+ and ProSight. As shown in Figure 4B, the number of matched fragment ions in pTop's individual identifications also tends to be larger than those in MS-Align+ and ProSight.

In summary, ProSight, MS-Align+, and pTop were used to search and analyze the three public data sets. As shown above, the identifications of pTop show more matched fragment ions both in the consensus and in the individual parts. The proteoforms identified in the human histone data set and the simulated data sets in the Supporting Information show that pTop can identify the proteoforms with more matched fragment ions. In the other two data sets (Figure S-6), pTop identified about 8% more proteins than ProSight and 4% to 46% more than MS-Align+.

Running Time Comparison. The preprocessing and search time of pTop were compared with those of ProSight and MS-Align+. As shown in Table 2, pParseTD ran about 3 times faster than MS-Deconv and about 10 times faster than Xtract, on average. For the searching speed, pTop was 90 to 118 times faster than MS-Align+ on the complex data sets (Table 2).

Table 2. Running Times of ProSight, MS-Align+, and pTop on Three Public Data Sets (Unit: Minutes)[a]

	ProSight		MS-Align+		pTop	
data sets	Xtract	Search	MS-Deconv	Search	pParseTD	Search
human[b]	269	263	194	4625	53	1725
E. coli	420	495	37	836	34	7
Autopilot	384	75	71	5756	14	63

[a]All of the running time tests were performed on the same PC (Intel (R) Core (TM) i7 CPU 870 at 2.93 GHz, Memory 12G). [b]The search time was recorded in the experiment to search against the human histone database by each software.

To compare the running time of pTop and MS-Align+ in the human histone data set, different numbers of modifications were used to search against a small database, which contained only the target proteins. The spectra exported from MS-Deconv were used for both MS-Align+ and pTop. The shift number of MS-Align+ was set as one or two, respectively. The variable modifications in pTop were set as nine modifications described in Table S-6. The maximum number of modifications allowed on each protein was one or two in pTop. TopPIC is an improved implementation of MS-Align+ (TopPIC: http://proteomics.informatics.iupui.edu/software/toppic/manual.html). We also tested it in this experiment.

As shown in Table 3, with the increase in the number of modifications, the running times of pTop and MS-Align+ increased. When the shift number of MS-Align+ was one or two, the running time of MS-Align+ was 48 and 69 times of that of pTop with one or two PTMs, respectively. If we do not know the target proteins in the sample or if a large protein database should be searched against, the running time of MS-Align+ will dramatically increase, reaching more than 20 times, and the memory required increases to 40G. It took 14.5 days for MS-Align+ to search the H2A data set against the whole human proteins database. pTop can complete such a search task within 1 day for 9 PTMs or 54 min for 4 PTMs. Compared with MS-Align+, TopPIC's running times have decreased a lot, as shown in Table 3. However, pTop was still about 15 to 17 times faster than TopPIC on this histone data set.

Table 3. Running Times of pTop, MS-Align+, and TopPIC (Unit: Minutes)[a]

	pTop		MS-Align+		TopPIC	
	1 PTM	2 PTMs	1 MOD[b]	2 MODs	1 MOD	2 MODs
H2A	17	19	772	1191	323	325
H2B	8	9	701	1351	436	412
H3	27	29	986	1383	235	206
H4	9	9	444	681	32	32
sum	61	67	2903	4606	1026	975

[a]All of the running time tests were performed on the same PC (Intel(R) Xeon(R) CPU E52670 at 2.6 GHz, Memory128 G). [b]The MOD here means the parameter "Shift Number" in MS-Align+ and TopPIC.

In summary, pTop took two procedures to decrease the search space of the proteoforms—first, it identified the sequence of the candidate proteins from the index of sequence tags, and, second, it characterized the combinatorial modifications using the mass difference between the theoretical and experimental precursors. As a result, the locations of the modifications would be considered only in a small search space. MS-Align+ allowed the shift number to be two with its search space expanded to $2 \times m \times n$, where m and n stand for the length of the mass lists of a spectrum and its candidate protein. The scale of candidate proteoforms in pTop depends on the number of combinatorial modifications under the constraint of the mass deviation of the precursor. For example, for a protein with a length of 100, the candidate proteoforms in MS-Align+ might be over 10 000, while it would be only several hundred in pTop. So the speedup of pTop, compared with MS-Align+, benefits mainly from the decrease in search space. In addition, the total memory requirement of MS-Align+ might go up to 40G when searching against the human protein database. However, the total memory of pTop is less than 1.5G, which can be conducted on a personal computer.

DISCUSSION

In order to improve the accuracy of protein precursor masses, we developed pParseTD to preprocess the raw data. We have shown its relative superiority in terms of RPR, RKR, or RIR on the human histone. Nevertheless, on the *E. coli* data, pParseTD's RPR in Table S-4 was similar to that of Xtract. One of the main reasons can be attributed to the difference of the sample complexity. The *E. coli* ribosome contains only about 56 proteins.[21] The relatively higher-quality spectra acquired in MS1 make both Xtract and pParseTD perform very well on RPR. pParseTD can adapt to deal with the cofragmented spectra by extracting multiple precursors, which leads to an improved RPR more significantly for the complex sample than the simple one. Besides the sample complexity, the performance comparisons on both RKR and RIR in Table S-5 were also related to the width of isolation window in MS1 for Autopilot and *E. coli* data sets, which is 15 and 20 m/z, respectively. A relatively larger window in *E. coli* data makes pParseTD perform better than Xtract to extract more correct

DOI: 10.1021/acs.analchem.5b03963
Anal. Chem. 2016, 88, 3082−3090

MS/MS peaks due to the fact that more peaks will be generated in MS/MS under a wider isolation window.

The averagine model[46] is adopted by most preprocessing algorithms either in bottom-up or top-down strategies. For the latter, there exists a phenomenon that we call "one-Dalton off," which means that there is frequently about one Dalton deviation between the correct precursor and the proteoform. In MS1 of the intact proteins, the monoisotopic peak can be calculated from the most abundant peak and its relative position in the theoretical isotopic distribution. However, the intensity difference between the most abundant peak and its neighbor one might be less than 1% if the mass of the precursor is larger than 10 kDa. Furthermore, the relative position of the most abundant peak in the theoretical isotopic distribution calculated by the averagine model may be different from the real isotopic distribution. For example, the most abundant peak in the theoretical isotopic distribution of the human histone H4 protein is the sixth peak, while that of the isotopic distribution calculated by the averagine model with a mass equal to 11 299 Da is the seventh one. All of these factors would possibly lead to a mass deviation of about ±1 Da from the real monoisotopic mass of the precursor ion. Therefore, we suggest that it is better to set the precursor tolerance to be more than 1.1 Da.

The main approach to obtaining the candidate proteins in pTop is to extract sequence tags from the MS/MS spectra. A large mass tolerance will be used for searching if there is no such adequate tag. In addition, the comparison in the complex data sets may be unfair for pTop, which took a strict limitation for mass deviation of precursors (such as 3.2 Da) since many PrSMs identified by ProSight have a larger mass deviation, such as −401.2 Da. Although the identifications of MS-Align+ have very small deviations (such as 20 ppm) from the observed precursor masses, there are still some mass deviations which are not explicitly interpreted (e.g., 1660.98 Da). For those large mass deviations, the corresponding identifications are in fact not verified. More accurate identifications may rely on the higher resolution of mass spectrometry and new bioinformatic approaches in the future.

Another issue is to identify the truncated proteoforms, whose N- or C-terminal peptide may be truncated. Current pTop 1.0 does not consider specifically about this kind of proteoform identification. If only one terminus is truncated, such as an N- or C-terminus, we can enlarge the search window to include the truncated version to be matched because the scoring function may still find it using the other terminal (not truncated) fragment ions which are not affected a lot. Nevertheless, we are ongoing to develop a new version to identify directly these truncated variants.

CONCLUSION

Bioinformatics in top-down proteomics has played an increasingly important role because the increased data sets and enhanced data quality demand for much more computational efforts. In this work, we developed pTop 1.0, aiming to improve the accuracy and efficiency of data analysis in TDP. Here, a model of multiple characteristics of the precursor ions was obtained by the online training of a machine-learning method—SVM. Compared with Xtract (in ProSight), although the number of exported precursors was reduced by 30%, the recalled precursor rate (RPR) of pParseTD (in pTop) was increased by 22% on a human histone data set. On the other hand, pTop can obtain the sequence of the candidate proteins and the combinatorial modifications by retrieving the indexes of sequence tags and modifications, which can reduce the search space of the proteoforms significantly. In complex data sets, pTop has proven to run about 1 or 2 orders of magnitude faster than MS-Align+.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.5b03963.

Additional information as noted in text; Figures S-1 to S-6 and Tables S-1 to S-8(PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: rxsun@ict.ac.cn.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This research was supported by a grant from National Basic Research Program of China (2013CB911203). The authors thank Kun He, Sheng-Bo Fan, Kun Zhang, Hao Yang, Jian-Qiang Wu, Hui-Jun Tu, and Ji-Li Yin from the Institute of Computing Technology; Zhe-Yi Liu, and Fang-Jun Wang from Dalian Institute of Chemical Physics, Chinese Academy of Sciences; and Meng-Qiu Dong from National Institute of Biological Sciences, Beijing, for valuable discussions.

REFERENCES

(1) Tran, J. C.; Zamdborg, L.; Ahlf, D. R.; Lee, J. E.; Catherman, A. D.; Durbin, K. R.; Tipton, J. D.; Vellaichamy, A.; Kellie, J. F.; Li, M.; Wu, C.; Sweet, S. M.; Early, B. P.; Siuti, N.; LeDuc, R. D.; Compton, P. D.; Thomas, P. M.; Kelleher, N. L. *Nature* **2011**, *480*, 254−258.

(2) Kelleher, N. L. *Anal. Chem.* **2004**, *76*, 196A−203A.

(3) Yates, J. R., 3rd; Kelleher, N. L. *Anal. Chem.* **2013**, *85*, 6151.

(4) Catherman, A. D.; Skinner, O. S.; Kelleher, N. L. *Biochem. Biophys. Res. Commun.* **2014**, *445*, 683−693.

(5) Catherman, A. D.; Durbin, K. R.; Ahlf, D. R.; Early, B. P.; Fellers, R. T.; Tran, J. C.; Thomas, P. M.; Kelleher, N. L. *Mol. Cell. Proteomics* **2013**, *12*, 3465−3473.

(6) Ahlf, D. R.; Thomas, P. M.; Kelleher, N. L. *Curr. Opin. Chem. Biol.* **2013**, *17*, 787−794.

(7) Arnaud, C. H. *Chem. Eng. News* **2013**, *91*, 11−17.

(8) Cui, W.; Rohrs, H. W.; Gross, M. L. *Analyst* **2011**, *136*, 3854−3864.

(9) Ntai, I.; Kim, K.; Fellers, R. T.; Skinner, O. S.; Smith, A. D. t.; Early, B. P.; Savaryn, J. P.; LeDuc, R. D.; Thomas, P. M.; Kelleher, N. L. *Anal. Chem.* **2014**, *86*, 4961−4968.

(10) Calligaris, D.; Villard, C.; Lafitte, D. *J. Proteomics* **2011**, *74*, 920−934.

(11) Gregorich, Z. R.; Ge, Y. *Proteomics* **2014**, *14*, 1195−1210.

(12) Peng, Y.; Gregorich, Z. R.; Valeja, S. G.; Zhang, H.; Cai, W. X.; Chen, Y. C.; Guner, H.; Chen, A. J.; Schwahn, D. J.; Hacker, T. A.; Liu, X. W.; Ge, Y. *Mol. Cell. Proteomics* **2014**, *13*, 2752−2764.

(13) Smith, L. M.; Kelleher, N. L.; et al. *Nat. Methods* **2013**, *10*, 186−187.

(14) Kelleher, N. L. *J. Am. Soc. Mass Spectrom.* **2012**, *23*, 1617−1624.

(15) Nesvizhskii, A. I.; Aebersold, R. *Mol. Cell. Proteomics* **2005**, *4*, 1419−1440.

(16) Siuti, N.; Kelleher, N. L. *Nat. Methods* **2007**, *4*, 817−821.

(17) Moradian, A.; Kalli, A.; Sweredoski, M. J.; Hess, S. *Proteomics* **2014**, *14*, 489−497.

(18) Olsen, J. V.; Macek, B.; Lange, O.; Makarov, A.; Horning, S.; Mann, M. *Nat. Methods* **2007**, *4*, 709−712.

DOI: 10.1021/acs.analchem.5b03963
Anal. Chem. 2016, 88, 3082−3090

(19) Syka, J. E.; Coon, J. J.; Schroeder, M. J.; Shabanowitz, J.; Hunt, D. F. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 9528−9533.
(20) Zubarev, R. A.; Horn, D. M.; Fridriksson, E. K.; Kelleher, N. L.; Kruger, N. A.; Lewis, M. A.; Carpenter, B. K.; McLafferty, F. W. *Anal. Chem.* **2000**, *72*, 563−573.
(21) Cannon, J. R.; Cammarata, M. B.; Robotham, S. A.; Cotham, V. C.; Shaw, J. B.; Fellers, R. T.; Early, B. P.; Thomas, P. M.; Kelleher, N. L.; Brodbelt, J. S. *Anal. Chem.* **2014**, *86*, 2185−2192.
(22) Meng, F.; Cargile, B. J.; Miller, L. M.; Forbes, A. J.; Johnson, J. R.; Kelleher, N. L. *Nat. Biotechnol.* **2001**, *19*, 952−957.
(23) Taylor, G. K.; Kim, Y. B.; Forbes, A. J.; Meng, F.; McCarthy, R.; Kelleher, N. L. *Anal. Chem.* **2003**, *75*, 4081−4086.
(24) LeDuc, R. D.; Taylor, G. K.; Kim, Y. B.; Januszyk, T. E.; Bynum, L. H.; Sola, J. V.; Garavelli, J. S.; Kelleher, N. L. *Nucleic Acids Res.* **2004**, *32*, W340−345.
(25) Zamdborg, L.; LeDuc, R. D.; Glowacz, K. J.; Kim, Y. B.; Viswanathan, V.; Spaulding, I. T.; Early, B. P.; Bluhm, E. J.; Babai, S.; Kelleher, N. L. *Nucleic Acids Res.* **2007**, *35*, W701−706.
(26) Fellers, R. T.; Greer, J. B.; Early, B. P.; Yu, X.; LeDuc, R. D.; Kelleher, N. L.; Thomas, P. M. *Proteomics* **2014**, *15*, 1235.
(27) Liu, X.; Sirotkin, Y.; Shen, Y.; Anderson, G.; Tsai, Y. S.; Ting, Y. S.; Goodlett, D. R.; Smith, R. D.; Bafna, V.; Pevzner, P. A. *Mol. Cell Proteomics* **2012**, *11*, M111 008524.
(28) Top Down Proteomics Virtual Issue. http://pubs.acs.org/page/vi/2013/topdown.html.
(29) Horn, D. M.; Zubarev, R. A.; McLafferty, F. W. *J. Am. Soc. Mass Spectrom.* **2000**, *11*, 320−332.
(30) Jaitly, N.; Mayampurath, A.; Littlefield, K.; Adkins, J. N.; Anderson, G. A.; Smith, R. D. *BMC Bioinf.* **2009**, *10*, 87.
(31) Mayampurath, A. M.; Jaitly, N.; Purvine, S. O.; Monroe, M. E.; Auberry, K. J.; Adkins, J. N.; Smith, R. D. *Bioinformatics* **2008**, *24*, 1021−1023.
(32) Guner, H.; Close, P. L.; Cai, W.; Zhang, H.; Peng, Y.; Gregorich, Z. R.; Ge, Y. *J. Am. Soc. Mass Spectrom.* **2014**, *25*, 464−470.
(33) Liu, X.; Inbar, Y.; Dorrestein, P. C.; Wynne, C.; Edwards, N.; Souda, P.; Whitelegge, J. P.; Bafna, V.; Pevzner, P. A. *Mol. Cell. Proteomics* **2010**, *9*, 2772−2782.
(34) Garcia, B. A. *J. Am. Soc. Mass Spectrom.* **2010**, *21*, 193−202.
(35) Perkel, J. M. *BioTechniques* **2012**, *53*, 75−78.
(36) Frank, A. M.; Pesavento, J. J.; Mizzen, C. A.; Kelleher, N. L.; Pevzner, P. A. *Anal. Chem.* **2008**, *80*, 2499−2505.
(37) Liu, X.; Hengel, S.; Wu, S.; Tolić, N.; Pasa-Tolić, L.; Pevzner, P. A. *J. Proteome Res.* **2013**, *12*, 5830−5838.
(38) Tsai, Y. S.; Scherl, A.; Shaw, J. L.; MacKay, C. L.; Shaffer, S. A.; Langridge-Smith, P. R.; Goodlett, D. R. *J. Am. Soc. Mass Spectrom.* **2009**, *20*, 2154−2166.
(39) Li, L.; Tian, Z. *Rapid Commun. Mass Spectrom.* **2013**, *27*, 1267−1277.
(40) Yuan, Z. F.; Liu, C.; Wang, H. P.; Sun, R. X.; Fu, Y.; Zhang, J. F.; Wang, L. H.; Chi, H.; Li, Y.; Xiu, L. Y.; Wang, W. P.; He, S. M. *Proteomics* **2012**, *12*, 226−235.
(41) Carvalho, P. C.; Xu, T.; Han, X.; Cociorva, D.; Barbosa, V. C.; Yates, J. R., 3rd. *Bioinformatics* **2009**, *25*, 2734−2736.
(42) Chi, H.; Chen, H.; He, K.; Wu, L.; Yang, B.; Sun, R. X.; Liu, J.; Zeng, W. F.; Song, C. Q.; He, S. M.; Dong, M. Q. *J. Proteome Res.* **2013**, *12*, 615−625.
(43) Tian, Z.; Tolic, N.; Zhao, R.; Moore, R. J.; Hengel, S. M.; Robinson, E. W.; Stenoien, D. L.; Wu, S.; Smith, R. D.; Pasa-Tolic, L. *Genome Biol.* **2012**, *13*, R86.
(44) Durbin, K. R.; Fellers, R. T.; Ntai, I.; Kelleher, N. L.; Compton, P. D. *Anal. Chem.* **2014**, *86*, 1485−1492.
(45) Rockwood, A. L.; Haimi, P. *J. Am. Soc. Mass Spectrom.* **2006**, *17*, 415−419.
(46) Senko, M. W.; Beu, S. C.; McLaffertycor, F. W. *J. Am. Soc. Mass Spectrom.* **1995**, *6*, 229−233.

DOI: 10.1021/acs.analchem.5b03963
Anal. Chem. 2016, 88, 3082−3090