Systematic Evaluation of Protein Sequence Filtering Algorithms for Proteoform Identification Using Top-Down Mass Spectrometry

Qiang Kou[1], Si Wu[2], and Xiaowen Liu[1,3,*]

[1]Department of BioHealth Informatics, Indiana University-Purdue University Indianapolis

[2]Department of Chemistry and Biochemistry, University of Oklahoma

[3]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine

Keywords: Top-down mass spectrometry, spectral identification, filtering algorithms

Significance of the study Identifying proteoforms with primary structural alterations is essential to understanding protein functions and related biological processes. In this study, we present new protein sequence filtering algorithms that outperform existing ones for top-down mass spectrometry-based proteoform identification. Combining the filtering algorithms and existing spectral alignment algorithms will significantly improve the sensitivity in proteoform identification and facilitate the studies of proteoforms with alterations.

Abstract

Complex proteoforms contain various primary structural alterations resulting from variations in genes, RNA, and proteins. Top-down mass spectrometry is commonly used for analyzing complex proteoforms because it provides whole sequence information of the proteoforms. Proteoform identification by top-down mass spectral database search is a challenging computational problem because the types and/or locations of some alterations in target proteoforms are in general unknown. Although spectral alignment and mass graph alignment algorithms have been proposed for identifying proteoforms with unknown alterations, they are extremely slow to align millions of spectra against tens of thousand protein sequences in high throughput proteome level analyses. Many software tools in this area combine efficient protein sequence filtering algorithms and spectral alignment algorithms to speed up database search. As a result, the performance of these tools heavily relies on the sensitivity and efficiency of their filtering algorithms. Here we propose two

Received: 08 16, 2017; Revised: 11 20, 2017; Accepted: 12 20, 2017

This is the author's manuscript of the article published in final edited form as:

Kou, Q., Wu, S. and Liu, X. (2018), Systematic Evaluation of Protein Sequence Filtering Algorithms for Proteoform Identification Using Top-Down Mass Spectrometry. Proteomics, 1700306. Accepted Author Manuscript. http://dx.doi.org/10.1002/pmic.201700306

efficient approximate spectrum filtering algorithms for proteoform identification. We evaluated the performances of the proposed algorithms and 4 existing ones on simulated and real top-down mass spectrometry data sets. Experiments showed that the proposed algorithms outperformed the existing ones for complex proteoform identification. In addition, combining the proposed filtering algorithms and mass graph alignment algorithms identified many proteoforms missed by ProSightPC in proteome-level proteoform analyses.

1 Introduction

Because of variations in genes, gene expression, and other biological processes, a gene may have various proteoforms [1], many of which contain multiple primary structural alterations (PSAs), such as amino acid substitutions, post-translational modifications (PTMs), and terminal truncations. PSAs play an essential role in many diseases such as heart failure [2] and age-dependent memory impairment [3]. Proteoform identification and PSA localization are essential to understanding proteoform functions in cellular processes. For example, the combinatorial patterns of PSAs in histone proteins determine their gene regulatory functions [4, 5].

Top-down mass spectrometry (MS) is more efficient than bottom-up MS in identifying modified proteoforms and combinatorial patterns of PSAs because it analyzes intact proteoforms instead of short protein fragments [6]. Database search is the dominant method for top-down MS-based proteoform identification, in which top-down tandem mass (MS/MS) spectra are searched against a protein sequence database or an annotated database for spectral identification [7-21]. A list of software tools for proteoform identification by top-down MS can be found in Table S1 in the supplementary material.

Most protein databases contain only a reference sequence for a gene or transcript even though the gene or transcript may have many proteoforms. The target proteoform may have various alterations compared with its corresponding database sequence [18]. We study four types of alterations: terminal truncations, fixed PTMs, variable PTMs, and unexpected alterations. A terminal truncation removes a prefix or suffix of a protein sequence. A fixed PTM specifies its mass shift and modified amino acids in a protein sequence. A variable PTM specifies its mass shift and a set of amino acids that may be modified. The mass shift and location of an unexpected alteration are unknown.

ProSightPC [7] is a commonly used software tool for top-down MS-based proteoform identification. By constructing a “shotgun annotated” proteoform database containing known modified proteoforms, it efficiently identifies proteoforms in the database. In the Delta-M mode of ProSightPC, the delta is the difference between the precursor mass of the query spectrum and the theoretical precursor mass of the target database sequence. When the allowed delta is large and a proteoform with unexpected alterations has a long unmodified N-terminal or C-terminal segment, mass spectra generated from the proteoform can be matched to its corresponding database sequence. In the sequence tag mode of ProSightPC, sequence tags are extracted from the query mass spectrum and searched against the proteoform database for proteoform identification.

Many software tools use spectral alignment to identify proteoforms with unexpected alterations [8, 12, 15-18]. Let *S* be a query mass spectrum generated from a proteoform with unexpected alterations and *P* the unmodified database protein sequence of the proteoform. The spectral alignment algorithm finds an optimal alignment between *S* and *P* by inserting into *P* mass shifts corresponding to the unexpected alterations in the blind mode. When the spectrum *S* contains enough fragment masses, the alignment algorithm is capable of identifying and characterizing the proteoform. MS-Align+ [12] and TopPIC [16] are commonly used tools for identifying proteoforms with unexpected alterations using top-down MS. In these tools, variable PTMs are treated as unexpected alterations, making them inefficient in identifying ultramodified proteoforms with many variable PTMs. To address this problem, several spectral alignment algorithms, such as MS-Align-E [15], MSPathFinder [21], pTop [17], TopMG [18] have been proposed to identify proteoforms with many variable PTMs.

There are two main steps in spectral alignment-based software tools for identifying proteoforms with variable PTMs and/or unexpected alterations by database search. First, a filtering algorithm is used to filter out most candidate protein sequences in the database for the query mass spectrum. Second, a spectral alignment algorithm is employed to align the mass spectrum against each remaining candidate protein sequence to find the best scoring proteoform spectrum-match (PrSM) [8]. It is extremely slow to align mass spectra against tens of thousands of database protein sequences [12]. Therefore, the filtering step is indispensable in proteome-level analyses. A filtering algorithm is *efficient* if it keeps the correct target protein sequence as a candidate for spectral alignment.

Most proteoform identification methods allow fixed PTMs and terminal truncations in the target proteoform. There are several scenarios for the other two types of alterations: (1) neither variable PTMs nor unexpected alterations are allowed in the target proteoform; (2) only variable PTMs are allowed; (3) only unexpected alterations are allowed; and (4) both variable PTMs and unexpected alterations are allowed. In the first scenario, a candidate protein sequence (may be truncated) is filtered out if its molecular mass does not match the precursor mass of the query spectrum. In the last three scenarios, the precursor mass of the query spectrum may be different from the molecular mass of its corresponding database sequence. For the second scenario, one filtering method is to check if the difference between the precursor mass and the molecular mass can be explained by a combination of variable PTMs. In this paper, we focus on filtering methods for the last three scenarios.

There are three main approaches for protein sequence filtering. In the first approach, a large error tolerance is allowed between the precursor mass of the query spectrum and the molecular mass of the candidate sequence [22]. In top-down MS, the method is employed in the Delta-M mode in ProSightPC [7]. However, when the error tolerance is very large, the filtering method reports many candidates, significantly increasing the running time of database search.

The second approach is based on sequence tags, which were proposed by Mann et al. in a pioneer work in 1994 [23]. In this approach, sequence tags are generated from the query spectrum and searched against the database to find hits, based on which top candidates are selected. Sequence tags and gapped sequence tags have been widely and successfully used for bottom-up

spectral interpretation [24-30]. In top-down MS, tag-based methods have been used in USTag [31], pTop [17], MSPathFinder [21], and the sequence tag mode in ProSightPC[7]. The accuracy of tag-based methods depends on whether the query spectrum contains consecutive fragment ions.

The third approach uses *unmodified protein fragments (UPFs)* and their matched fragment masses in the query spectrum to filter proteins [12, 16]. The idea is to find a mass shift for the fragment masses in the query spectrum such that many shifted fragment masses are explained by the unmodified target protein sequence. This method is computationally intensive. Fortunately, index-based algorithms [32-34] have been proposed to partially solve the problem. In top-down MS, UPF-based methods have been used in MS-Align+ [12] and TopPIC [16] and achieved satisfactory performance in identifying unexpected alterations. The three filtering approaches can be combined to improve filtering efficiency. For example, proteins can be filtered by combining a large error tolerance for the precursor mass and sequence tags extracted from the query spectrum.

The three filtering approaches were designed to identify proteoforms with a limited number (1 or 2 in most cases) of unexpected alterations. These methods may fail to keep the target database protein sequence in filtration when the target proteoform contains more than 2 variable PTMs and/or unexpected alterations.

In this paper, we propose *two approximate spectrum filtering (ASF)* algorithms for identifying complex proteoforms with variable PTMs and those with both variable PTMs and unexpected alterations. Let *F* be the target proteoform and *F'* a proteoform obtained from *F* by removing *h* variable PTMs. In the ASF algorithms, the query spectrum is transformed into an approximate spectrum of *F'*, which is searched against database sequences to find candidate proteins. After the transformation, the number of variable PTMs in the target proteoform is reduced by *h* (Fig. 1), significantly increasing filtering efficiency.

We evaluated the ASF algorithms and 4 existing ones for protein filtration in top-down MS database search. Experiments on simulated data showed that the ASF algorithms outperformed the existing ones for complex proteoform identification. Combining the ASF and mass graph alignment algorithms [18] identified many phosphorylated proteoforms missed by ProSightPC from a top-down MS data set of breast cancer xenograft samples.

2 Methods

A top-down MS/MS spectrum contains a list of peaks, each of which is represented as (*m/z, intensity*), where *m/z* and *intensity* are the mass-to-charge ratio and abundance of its corresponding fragment ion, respectively. The precursor mass of the MS/MS spectrum measures the molecular mass of the target proteoform. The first step in top-down spectral interpretation is usually spectral deconvolution [21, 35-41], which converts fragment ion peaks of various charge states and isotopomers into neutral monoisotopic fragment masses. A list of software tools for top-down spectral deconvolution can be found in Table S2 in the supplementary material. MS-Deconv [39] was used in the experiments for spectral deconvolution. In MS-Deconv, candidate isotopomer envelopes, each of which contains peaks from the same fragment ion with the same charge state, are first obtained by using the theoretical intensity distributions of these peaks, and are then selected by a

dynamic programming algorithm. Finally, a neutral monoisotopic mass is computed for each selected isotopomer envelope. MS-Deconv often significantly simplifies top-down MS/MS spectra and convert a complex spectrum with thousands of peaks into a deconvoluted one with dozens or hundreds of fragment masses. We assume that the query spectrum is a deconvoluted top-down MS/MS spectrum in database search.

In the ASF filtering algorithms, approximate spectra are first generated from the query spectrum and then searched against the protein database using the methods proposed in UPF-based filtering algorithms. We first review UPF-based filtering algorithms and then describe the ASF filtering algorithms.

2.1 UPF-based filtering algorithms

We introduce some notations for describing UPF-based filtering algorithms. Let *mass(a)* be the residue mass of an amino acid *a*. The residue mass of a protein sequence $P = a_1 a_2 ... a_n$ is the sum of the residues masses of its amino acids, that is, $\sum_{k=1}^{n} mass(a_k)$. The residue mass of the length-*i* prefix $a_1 a_2 ... a_n$ is a prefix residue mass of *P*, denoted by p_i. The residue mass of the length-*i* suffix $a_{n-i+1} ... a_n$ is a suffix residue mass of *P*, denoted by s_i. Specifically, the residue masses of the empty prefix and the empty suffix are 0, that is, $p_0 = 0$ and $s_0 = 0$. We denote the set of all prefix residue masses of *P* as $P_{pre} = \{p_0, p_1, ..., p_n\}$ and the set of all suffix residue masses of *P* as $P_{suf} = \{s_0, s_1, ..., s_n\}$.

Let *S* be a deconvoluted top-down MS/MS spectrum with a precursor mass M. The set of deconvoluted neutral fragment masses of S are converted into a set of possible prefix (suffix) residue masses corresponding to the masses of proteoform prefixes (suffixes). When S is a collision-induced dissociation (CID) spectrum, both the prefix residue mass set and the suffix residue mass set contain the following two masses: 0 and *M* - mass(H_2O), where mass(H_2O) is the mass of a water molecule. In addition, for each fragment mass *x*, two masses *x* and *M* - *x* are added to the prefix residue mass set, and two masses *x* - mass(H_2O) and *M* - *x* - mass(H_2O) are added to the suffix residue mass set. The mass of a water molecule is deducted from *x* for suffix residue masses because the mass difference between a neutral y-ion fragment mass and its corresponding suffix residue mass is mass(H_2O). The sets of fragment masses, prefix residue masses, and suffix residue masses of spectrum *S* are denoted as S_{fra}, S_{pre}, and S_{suf}, respectively. For example, when *S* is a CID spectrum with a precursor mass 302.17 Da and two neutral fragment masses 71.04 Da and 174.11 Da, the mass 0 and *M* - mass(H_2O) = 284.17 are added into S_{pre} and S_{suf}. S_{pre} = {0, 71.04, 128.06, 174.11, 231.13, 284.17} after the masses *x* and *M* - *x* for fragment masses *x* are added; S_{suf} = {0, 53.04, 110.06, 156.11, 213.13, 284.17} after the masses *x* - mass(H_2O) and *M* - *x* - mass(H_2O) for *x* are added. Similarly, we use the most commonly observed fragment ion types to convert other types of deconvoluted spectra into prefix (suffix) residue masses. For example, when we choose c, z-dot, and z-prime ions as the most commonly observed ones in electron-transfer dissociation (ETD) spectra, each fragment mass in the deconvoluted spectrum is converted to three possible prefix residue

masses based on the mass differences between the neutral prefix residue mass and its corresponding c, z-dot and z-prime fragment masses.

Two UPF-based filtering methods are implemented in TopPIC [16]. The first method is based on diagonal scores defined below. Let *A, B* be two set of masses. The mass counting score of *A* and *B* is the number of masses in *A* that match masses in *B* (within an error tolerance), denoted by *C(A, B).* Let *shift(A, d)* be the set of masses generated by adding a shift d to each mass in *A*. The diagonal score of *A* and *B* is the maximum mass counting score of *A* and *B* among all shift values (Fig. 2(a)), denoted by $D(A, B) = max_d C(shift(A, d), B)$. Let *P* be an unmodified protein sequence and *F* a modified form of *P* with truncations and PTMs. A high diagonal score between P_{pre} and F_{pre} means that *F* contains a long unmodified fragment. For example, the proteoform T[Ph]IDEST[Ph]R in Fig. 2(a) contains an unmodified fragment IDES. When a CID spectrum of T[Ph]IDEST[Ph]R contains peaks of the b-ions $b_1, b_2, ..., b_5$, the diagonal score between the prefix residue masses of PEPTIDESTRING and those of the spectrum is at least 5. In the first method, the similarity score between a database protein sequence *P* and a deconvoluted spectrum *S* is defined as $D(P_{pre}, S_{pre})$.

The second method is based on restricted diagonal scores. The restricted diagonal score of *A* and *B* is the maximum mass counting score among all non-positive shifts whose absolute values equal a mass in *A* (Fig. 2(b)), denoted by $R(A,B) = \max_{d \in A} C(shift(A,-d),B)$. For example, when *A* is the set of prefix residue masses {0, 97.05, 226.09} of the peptide PE, $R(A, B) = max\{C(shift(A, 0), B), C(shift(A, -97.05), B), C(shift(A, -226.09), B)\}$. A high restricted diagonal score between P_{pre} and F_{pre} means that *F* contains a long unmodified prefix that is a substring of *P*. For example, the proteoform TIDEST[Ph]R in Fig. 2(b) contains an unmodified prefix TIDES that is a substring of PEPTIDESTRING. In contrast, the restricted diagonal score between the prefix residue masses of T[Ph]IDEST[Ph]R and those of PEPTIDESTRING is 1 because T[Ph]IDEST[Ph]R does not have a long unmodified prefix. Similarly, a high restricted diagonal score between P_{suf} and F_{suf} means that *F* contains a long unmodified suffix that is a substring of *P*. In the second method, the similarity score between a protein sequence *P* and a deconvoluted spectrum *S* is defined as $R(P_{pre}, S_{pre}) + R(P_{suf}, S_{suf})$, which is determined by the unmodified prefix and suffix of the target proteoform. Different from the diagonal score, only a small number of mass shifts are considered to compute a restricted diagonal score. As a result, the chance that a random spectrum protein pair has a high restricted diagonal score is significantly reduced compared with the diagonal score. However, when the target proteoform has two modifications: one at the N-terminus and the other at the C-terminus, using the restricted diagonal score may fail to retain the target database protein sequence. The second method is efficient for identifying proteoforms with a long unmodified prefix or suffix.

In the two filtering methods, the two similarity scores are used to rank proteins in the database, and the top *t* proteins are reported as filtering results. The scores are computed using index-based algorithms [32]. The two methods are called UPF-DIAGONAL (the diagonal score) and UPF-RESTRICT (the restricted diagonal score), respectively.

2.2 ASF algorithms

In bottom-up MS, variable PTMs are often incorporated into database peptides to identify modified peptides. However, this approach is inefficient for top-down MS (see Section Discussion). In the proposed ASF algorithms, we incorporate variable PTMs into the query spectrum to improve the efficiency and sensitivity of protein filtration.

We use phosphorylation as an example to explain how to generate an approximate spectrum. Let δ be the mass shift of phosphorylation. Let $P = a_1 \ldots a_i \ldots a_n$ be an unmodified protein sequence (may be truncated) and *F* a modified form of *P* with one phosphorylation site on the amino acid a_i. The theoretical prefix residue mass spectrum $P_{pre} = \{p_0, p_1, \ldots, p_i, p_{i+1}, \ldots, p_n\}$ contains all prefix residue masses of *P* and the theoretical spectrum F_{pre} contains all prefix residue masses of *F*, that is, $F_{pre} = \{p_0, p_1, \ldots, p_i + \delta, p_{i+1} + \delta, \ldots, p_n + \delta\}$. We can convert F_{pre} into P_{pre} by deducting δ from the prefix residue masses $p_i + \delta, p_{i+1} + \delta, \ldots, p_n + \delta$.

Let S_{pre} be a prefix residue mass spectrum generated from an experimental spectrum of *F*. The precursor mass of the experimental spectrum is *M*. The spectrum S_{pre} is similar to F_{pre}, but has missing and noise peaks. To simplify the analysis, we assume that S_{pre} is a perfect spectrum, that is, $S_{pre} = F_{pre} = \{p_0, p_1, \ldots, p_i + \delta, p_{i+1} + \delta, \ldots, p_n + \delta\}$. In the ASF method, we try to convert S_{pre} into an approximate spectrum of P_{pre} with limited information (Fig. 1): it is known that the target proteoform contains a phosphorylation, but the target protein sequence and the location of the phosphorylation site are unknown.

Because the modification site is unknown, we give *k* guesses for the prefix residue mass p_i, the smallest prefix residue mass with the modification, and hope that one of the guesses is similar to p_i. The mass $p_n + \delta$ in S_{pre} is the residue mass of the target proteoform, which equals *M* - mass(H_2O). We divide the mass $p_n + \delta$ into *k* intervals $(0, l], (l, 2l], \ldots, ((k-1)l, kl]$ each with the same length $l = \frac{p_n + \delta}{k}$. The *k* centers of the intervals are the guessed values for p_i. For example, when $p_n + \delta =$ 5000 Da and $k = 2$, the two intervals are (0, 2500] and (2500, 5000], and the two centers are 1250 and 3750.

For each guessed prefix residue mass *q*, we convert S_{pre} into a spectrum conv(S_{pre}, *q*) by deducting δ from all masses in S_{pre} that are no less than *q*. In Fig. 1, the guessed prefix residue mass is 200 Da and all masses no less than 200 Da are shifted to the left by 79.97 Da. When $q < p_i$, all masses in the mass intervals (*0, q*) and [p_i, $p_n+\delta$] are correctly converted into their corresponding masses in P_{pre}, and all masses in the mass interval [*q*, p_i) are not correctly converted. In Fig. 1, peaks in the mass intervals (0, 200) and [546.14, 799.29] are correctly converted into peaks of TYPDSRP, but the left most peak in the box is not correctly converted. The ratio between the length of the interval [*q*, p_i) and $p_n + \delta$ is called the conversion error ratio of conv(S_{pre}, *q*). When $q > p_i$, all masses in the mass intervals (0, p_i) and [*q*, $p_n + \delta$] are correctly converted into their corresponding masses in P_{pre}, and all masses in the mass interval [p_i, *q*) are not correctly converted. The conversion error ratio of conv(S_{pre}, *q*) is the ratio between the length of the interval [p_i, *q*) and $p_n + \delta$. The distance between

p_i and the best guessed value q^* is no larger than $\frac{1}{2k}$, and the conversion error ratio of conv(S_{pre}, q^*) is no larger than $\frac{1}{2k}$. When k is large, conv(S_{pre}, q^*) is almost the same as P_{pre} and is called an *approximate prefix residue mass spectrum* of P. In practice, although S_{pre} has missing and noise peaks, it is converted into an approximate prefix residue mass spectrum of P using the same method. The above method is used to generated approximate suffix residue mass spectra as well.

Next, we extend the method to generate approximate spectra for proteoforms with $g > 1$ variable PTM sites. When the target proteoform F is ultramodified and the number g is large, it is impractical to enumerate all approximate spectra with g PTM sites. Let F' be a proteoform that is obtained from F by removing h variable PTM sites. By using h ($h < g$) variable PTM sites in spectral conversion, we generate an approximate spectrum of F' from S_{pre}. Although the resulting spectrum is not an approximate spectrum of the protein sequence P, it is more similar to the theoretical spectrum of P compared with S_{pre}. We treat the remaining $g - h$ PTM sites in F' as unexpected PTMs. Note that h is a user-specified parameter and not related to the number of PTM sites in the target proteoform.

To generate approximate spectra, we first choose h interval centers (each of the k centers can be chosen multiple times) as the guessed values of the prefix residue masses corresponding to the h PTM sites, then enumerate all possible combinations of the types of variable PTMs on the sites. For each configuration of h guessed prefix residue masses and guessed PTM types, we convert the spectrum S_{pre} into an approximation spectrum. The total number of configurations is proportional to $(kf)^h$. The UPF-RESTRICT and UPF-DIAGONAL methods are employed to search these approximate spectra against the protein database to find candidate proteins. The ASF method coupled with UPF-RESTRICT is called the ASF-RESTRICT algorithm (Fig. S1). Detailed steps for Step 4 in the algorithm is given in Fig. S2 in the supplementary material. To couple the ASF method with UPF-DIAGONAL, we replace the UPF-RESTRICT algorithm with the UPF-DIAGONAL algorithm in Step 5 of the ASF-RESTRICT algorithm. The ASF method with the UPF-DIAGONAL algorithm is referred to as the ASF-DIAGONAL algorithm.

To guarantee the efficiency of the method, the values of k, f and h need to be small. In the experiments, $k=3$ was chosen based on the evaluation of speed and sensitivity of the ASF algorithms with various settings of k (see Section Parameter settings), and h was set as 1 or 2. The number f of variable PTM types is a parameter specified by the user.

3 Results

3.1 Data sets

Four top-down MS data sets were used in this study: the first was generated from *Escherichia coli* (EC) K-12 MG1655, the second from purified human histone H3 protein, the third from purified human histone H4 protein, and the fourth from breast tumor xenograft samples.

The EC data set was obtained using a liquid chromatography system coupled with an LTQ Orbitrap Velos mass spectrometer (Thermo Scientific, Waltham, MA). MS and MS/MS spectra were collected at a 60000 resolution. The top 4 ions in each MS spectrum were selected for MS/MS analysis and the alternating fragmentation mode was used. In total, 2027 CID and 2027 ETD top-down MS/MS spectra were collected [16].

The histone H3 data set [42] was obtained using an LTQ Orbitrap Velos mass spectrometer (Thermo Scientific, Waltham, MA). Core histones were separated in the first dimension using a Jupiter C5 column and further separated in the second dimension by a weak cation exchange hydrophilic interaction LC (WCX-HILIC) using a PolyCAT A column. All acquisitions were performed with a 60000 resolution. In total, 3462 CID and 3462 ETD top-down MS/MS spectra were collected.

The histone H4 data set [15] was generated using an LTQ Orbitrap Velos mass spectrometer (Thermo Scientific, Waltham, MA). Core histones were separated by a 2-dimensional reversed-phase and hydrophilic interaction liquid chromatography (RP-HILIC) system where the histone H4 protein was isolated in the first dimension. With a resolution of 60000, a total of top-down MS/MS 1626 CID and 1626 ETD spectra were acquired.

The breast tumor xenograft data set [43] was generated using an Orbitrap Elite mass spectrometer (Thermo Scientific, Waltham, MA). Cryopulverization of the tumor xenografts was performed using the standard CPTAC protocols [44]. A basal-like (WHIM2) breast cancer sample and a luminal B (WHIM16) breast cancer sample [45, 46] were used for the experiments. Protein separation was achieved using a commercial GELFREE 8100 fractionation system (Expedeon, Cambridge, UK). With a resolution of 60000, a total of 51474 and 50372 higher-energy collisional dissociation (HCD) top-down MS/MS spectra were collected from the WHIM2 and WHIM16 samples respectively.

3.2 Simulated data set

To evaluate the accuracy and speed of the filtering algorithms, a test data set of PrSMs with mutations (treated as PTMs) was generated from the EC data set. The proteome database of *Escherichia coli* K-12 MG1655 was downloaded from the UniProt database [47] (version Sept 12, 2016, 4306 entries) and concatenated with a shuffled decoy database of the same size. The 4054 top-down MS/MS spectra were deconvoluted by MS-Deconv [39] and then searched against the target-decoy concatenated EC proteome database using TopPIC [16]. Parameter settings of TopPIC are given in Table S3 in the supplementary material. A total of 874 PrSMs without PTMs (529 from CID and 345 from ETD) were identified with a 1% spectrum-level false discovery rate (FDR). The 874 PrSMs can be found in Table S4 in the supplementary material, and the histogram of the lengths of the identified proteoforms is given in Fig. S3 in the supplementary material.

For each identified PrSM between a spectrum S and a protein sequence P with a score x, we used the generating function method [48, 49] to compute the conditional spectral probability that the similarity score between the spectrum S and a random protein sequence is no less than x on the condition that the molecular mass of the random protein matches the precursor mass of S. In the generating function method, a dynamic programming algorithm is employed to efficiently and

accurately compute the distribution of the similarity scores between the spectrum *S* and random proteins as well as the conditional spectral probability. The histogram of the conditional spectral probabilities of the identified PrSMs is given in Fig. S4 in the supplementary material.

The 874 PrSMs without PTMs were used to generate test PrSMs with random mutations. Let (*P, S*) be a PrSM between a spectrum *S* and a protein sequence *P* without PTMs. We randomly select an amino acid in *P*, then replace it with a random amino acid, resulting in a protein sequence *P'* with a mutation. The mass difference between the original amino acid and the new one is required to be larger than 5 Da. In addition, a random sequence with no more than 20 amino acids is appended to the N-terminus of *P'* and another random sequence with no more than 20 amino acids to the C-terminus of *P'*. The PrSM between the resulting sequence and S contains a PTM (mutation), an N-terminal truncation, and a C-terminal truncation. Using this method, a total of 13110 test PrSMs (15 test PrSMs for each of the 874 PrSMs: 5 without terminal truncation, 5 with only an N- or C-terminal truncation, and 5 with both N- and C-terminal truncations) were generated. In addition, PrSMs with 2, 3, 4, 5 mutations were generated using a similar method. When two or more PTMs (mutations) were added to a protein sequence, the random mutations were chosen independently and were different in most cases. A total of 65550 PrSMs (13110 for each setting of the mutation numbers 1, 2, 3, 4, 5) were generated. All the experiments on the simulated data set were performed on a desktop with an Intel Core i7-3770 Quad-Core 3.4 GHz CPU and 16 GB memory.

3.3 Parameter settings

We tested the ASF-RESTRICT and ASF-DIAGONAL algorithms with various settings of the parameters *k* and *h* on the simulated PrSMs with 5 PTMs. The error tolerance for computing diagonal scores and restricted diagonal scores was 15 ppm. For each test PrSM with a mutated protein sequence *P'* and a spectrum *S*, we replaced the unmodified protein sequence of *P'* in the EC proteome database with *P'*, then used the ASF algorithms to search *S* against the proteome database, and finally reported t = 20 candidate proteins. If the 20 candidate proteins contain protein *P'*, we say the filtration is efficient. The efficiency rate of the filtering algorithm is the ratio between the number of PrSMs with efficient filtration and the total number of test PrSMs.

The efficiency rates and average running times (per spectrum) of the ASF algorithms with various settings for k = 2, 3, 4, 5, 6 and h = 1, 2 are shown in Fig. 3 and Fig. S5 in the supplementary material. Removing two modification sites from the query spectrum (h = 2) achieved marginal improvement in the efficiency rate compared with removing one modification site (h = 1). However, the average running time of ASF-RESTRICT and ASF-DIAGONAL with h = 2 was more than 10 times slower than those with h = 1. When k increases, the efficiency rate increases, but the increase rate becomes less significant. In the ASF-based methods, each approximate spectrum is searched against the database sequentially, and the memory usage of the algorithms remains the same when the parameter settings of h and k increase and the number of generated approximate spectra increases. The memory usage of ASF-RESTRICT and ASF-DIAGONAL was less than 4 GB.

3.4 Evaluation on filtration efficiency

Two sequence tag-based filtering methods were compared with the UPF and ASF-based methods on the simulated PrSMs. The first method, which was employed in MS-Align+Tag (http://bioinf.spbau.ru/proteomics/ms-align-plus-tag), uses the long tag strategy. Long tags are first extracted from the query spectrum, then all length *l* (*l* = 4 in the experiments) substrings of the long tags are reported for protein sequence filtration. The second method, which is a part of MSPathFinder [21], extracts from the query spectrum all sequence tags with a length *l* between the minimum length l_{min} and the maximum length l_{max}, that is, $l_{\min} \le l \le l_{\max}$. In the experiment, l_{min} = 5 and l_{max} = 8. The two methods are called TAG-LONG (with the long tag strategy) and TAG-VAR (with tags of various lengths), respectively. Detailed description of the two tag-based methods can be found in the supplementary material.

We tested the TAG-LONG, TAG-VAR, UPF-RESTRICT, UPF-DIAGONAL, ASF-RESTRICT and ASF-DIAGONAL algorithms on the simulated PrSMs with 5 PTMs. Parameter settings of the algorithms are given in Table S5 in the supplementary material. The ASF-DIAGONAL method achieved the best filtration efficiency rate 82.4%, while the filtration efficiency rates of the tag-based methods were below 40% and those of the UPF-based method were below 70% (Fig. S7 in the supplementary material). The ASF-DIAGONAL algorithm missed 528, 253, and 794 PrSMs efficiently filtered by UPF-RESTRICT, UPF-DIAGONAL, and ASF-RESTRICT, respectively (Fig. S8 in the supplementary material).

The efficiency rates of the filtering algorithms are related to the conditional spectral probabilities of test PrSMs (Fig. 4). Most PrSMs with a conditional spectral probability $\ge 10^{-30}$ have less than 30 matched masses, and protein sequence filtering for these PrSMs is more challenging than those with many matches masses. For PrSMs with a conditional spectral probability between 10^{-20} and 10^{-30}, the efficiency rate of ASF-DIAGONAL was higher than 85%. For PrSMs with a conditional spectral probability between 10^{-10} and 10^{-20}, the efficiency rate of the ASF-DIAGONAL algorithm was still higher than 50%. In addition, the filtration efficiency rates of ASF-based algorithms were similar on CID and ETD spectra (Fig. S9 in the supplementary material).

The filtration efficiency rates of the algorithms for the simulated test PrSMs with 1, 2, 3, and 4 PTMs are shown in Fig. S10-S13 in the supplementary material. Because ASF-RESTRICT and ASF-DIAGONAL are designed for identifying proteoforms with multiple PTMs, they were not tested on the PrSMs with 1 PTM. ASF-RESTRICT outperformed the other algorithms on the test PrSMs with 2 or 3 PTMs, and ASF-DIAGONAL obtained the best performance on the test PrSMs with 4 or 5 PTMs. The main reason is that ASF-RESTRICT and ASF-DIAGONAL have complementary strengths in protein sequence filtration. When the proteoform that corresponds to the approximate spectrum contains only a small number of PTMs, it is highly possible that the proteoform has a long unmodified N-terminal or C-terminal segment. Compared with ASF-DIAGONAL, ASF-RESTRICT is more efficient for identifying this type of proteoforms. ASF-DIAGONAL is more powerful than ASF-RESTRICT when the proteoform contains a long unmodified internal segment. The experimental results show that combining the two methods can improve filtration efficiency.

The average running time of ASF-DIAGONAL (10.9 seconds) for one test PrSM was about 8 times of TAG-LONG (1.34 seconds) and TAG-VAR (1.35 seconds) and 13 times of UPF-DIAGONAL (0.85 seconds). Although ASF-DIAGONAL is slower than other filtering methods, its running time is still acceptable because the running time is similar to that of spectral alignment algorithms. The running time for aligning a mass spectrum with 20 candidate protein sequences is usually more than 20 seconds.

To test the filtering algorithms on large protein databases, we concatenated the EC proteoform database with the human proteome database downloaded from the UniProt database [47] (version Jul 9, 2016, 20191 entries). The concatenated database contained 24497 proteins. The filtration efficiency rates of ASF-RESTRICT and ASF-DIAGONAL were 61.6% and 70.6%, respectively, while those of the other four algorithms were below 55% (Fig. S14 in the supplementary material).

3.5 Evaluation on the histone data sets

The two human histone protein data sets were used to evaluate the filtering methods for identifying proteoforms with multiple PTMs. All the experiments on the histone data sets were performed on the same desktop used for the simulated data analyses. All the spectra of the histone H3 and H4 data sets were deconvoluted using MS-Deconv [39]. TopMG [18] was employed to align the histone H3 and H4 spectra against their corresponding histone H3 and H4 protein sequences. Five PTMs: acetylation, methylation, dimethylation, trimethylation, phosphorylation (Table S6 in the supplementary material) were used as variable PTMs in proteoform identification. Other parameter settings used in TopMG are given in Table S7 in the supplementary material. TopMG identified 3205 and 1087 PrSMs with at least 10 matched fragment ions from the histone H3 and H4 data sets, respectively (Table S8 and S9 in the supplementary material).

The tag-based, UPF-based, and ASF algorithms were tested on these identified PrSMs. For each identified PrSM of protein *P* and spectrum *S*, the filtering algorithm used the spectrum *S* to filter the UniProt human proteome database (version Jul 9, 2016, 20191 entries) and reported 20 top candidate protein sequences. If the 20 protein sequences contain the target protein *P* (histone H3 or H4), the filtration is efficient. The five PTMs used in proteoform identification were treated as variable PTMs in the ASF algorithms, and parameter settings of the algorithms are provided in Table S5 in the supplementary material.

The filtration efficiency rates of the 6 filtering methods for the histone H3 and H4 PrSMs are summarized in Table 1. The filtration efficiency rates of the two tag-based methods were not as high as the UPF and ASF based methods. The main reason is that many spectra in the test PrSMs do not contain long consecutive fragment ions. The filtration efficiency rates of UPF-RESTRICT and ASFRESTRICT were the highest among the 6 methods. Most of the histone H3 and H4 proteoforms have no more than 4 PTMs (Fig. S15(b) and Fig. S16(b) in the supplementary material), and most PTM sites on the histone H3 and H4 proteins lie in a short region near the N-terminus and can be treated as one large unexpected mass shift in protein filtering. UPF-RESTRICT and ASF-RESTRICT are efficient in filtering proteins for this type of spectra. As a result, ASF-RESTRICT outperformed ASF-DIAGONAL on the histone data sets. Compared with UPF-RESTRICT, ASFRESTRICT improved the efficiency rate by about 9.7% for the histone H3 PrSMs and 2.6% for the histone H4 PrSMs. ASF-

RESTRICT efficiently filtered 334 histone H3 PrSMs missed by UPFRESTRICT and 1094 histone H3 PrSMs missed by ASF-DIAGONAL (Fig. S17(a)). Similarly, ASF-RESTRICT outperformed ASF-DIAGONAL and UPF-RESTRICT on the histone H4 PrSMs (Fig. S17(b)). The Venn diagrams for the comparison of ASF-RESTRICT, ASF-DIAGONAL, TAG-LONG, and TAG-VAR can be found in Fig. S18 in the supplementary material. Compared with UPF-RESTRICT, ASF-RESTRICT achieved a better improvement on the histone H3 data set than the histone H4 data set. The main reason is that the quality of the histone H3 PrSMs is not as good as that of the histone H4 PrSMs. While 86.0% of the histone H3 PrSMs contain ≤ 25 matched fragment ions, only 29.7% of the histone H4 PrSMs contain ≤ 25 matched fragment ions (Fig. S15(a) and S16(a) in the supplementary material). Most of the PrSMs with ≤ 25 matched fragment ions have a relatively large conditional spectral probability. Compared with the UPF-based methods, the ASF algorithms achieve a better improvement in the filtration efficiency for PrSMs with large conditional spectral probabilities than those with very small ones (Fig. 4).

A total of 892 histone H3 PrSMs and 7 histone H4 PrSMs were missed by ASF-RESTRICT. The main reasons for inefficient filtration of these PrSMs are: (1) some PrSMs are of low quality and (2) some contain many PTM sites. Of the 899 histone PrSMs (892 histone H3 and 7 histone H4 PrSMs), 576 (64.1%) contain no more than 15 matched fragment ions. Of the other 323 PrSMs, 294 (91.0%) contain at least 4 variable PTM sites. Of the 29 remaining PrSMs, 28 have less than 22 matched fragment ions but more than 220 deconvoluted peaks and 1 has 125 deconvoluted peaks with 17 matched fragment ions, showing the low quality of the PrSMs.

The speed of the ASF algorithms is much slower than the other filtering methods. For the histone H3 data set, the running time of ASF-RESTRICT was about 11 times of UPF-RESTRICT, and the running time of ASF-DIAGONAL was about 11 times of ASF-RESTRICT and 130 times of UPF-RESTRICT. In practice, the ASF-based algorithms can be combined with other methods to speed up protein sequence filtration: fast filtering methods are used in the first round of spectral identification, and the ASF-based algorithms are employed to identify spectra that are elusive for the fast methods.

3.6 Phosphorylated proteoforms identified from the xenograft data set

The ASF algorithms were combined with TopMG [18] for proteome-wide complex proteoform identification. In the combined method, ASF-RESTRICT and ASF-DIAGONAL were employed to report top 20 candidate proteins separately for each query spectrum. The resulting proteins were aligned with the query spectrum using TopMG to find the best PrSM. We compared the performances of ProSightPC [7] and TopMG coupled with the ASF algorithms for identifying phosphorylated proteoforms on the breast cancer xenograft data set.

All the mass spectra from the WHIM2 and WHIM16 samples were deconvoluted by MS-Deconv [39]. Because the xenograft samples contain both mouse and human proteins, a multi-step database search approach was used for proteoform identification. While TopMG coupled with the ASF methods was used to identify phosphorylated proteoforms, TopPIC [16] was used to identify proteoforms without variable PTMs. The experiments were performed on a node with two 12-core Intel Xeon E5-2680 v3 CPUs and 256 GB memory on Carbonate, a parallel computing system at Indiana University. A total of 12 threads were used in the analysis. The running time for analyzing all

the spectra was about 63 hours (3 hours for TopPIC and 60 hours for TopMG), of which 30 hours were used by the ASF algorithms. When multiple threads are used, the memory usage of the ASF algorithms is proportional to the number of threads. The maximum memory usage for analyzing the xenograft data set was 48 GB (4 GB for each thread).

Proteoforms identified by ProSightPC were obtained from a previous study [43], in which a customized version of cRAWler was used for spectral deconvolution and a five step database search was performed for proteoform identification. The third and fourth steps were to identify proteoforms with sample specific mutations and splicing events; the fifth step was to identify proteoforms with unexpected alterations. Because the last three steps were not designed to identify proteoforms with variable PTMs, we focused on only proteoforms identified in the first two steps.

Mouse proteoforms In the first step of the ProSightPC analysis, the absolute mass mode was used to search all the deconvoluted spectra against a mouse proteoform database including proteoforms with PTMs, which was built based on the UniProt mouse proteome database (version May 2014) and its annotations. The error tolerances for precursor and fragment masses were set as 2.2 Da and 10 ppm, respectively. With a *p*-value cutoff 10^{-10}, this step reported 648 proteoforms from 54 proteins, including 41 proteoforms without PTMs (N-terminal acetylation is allowed) and 24 phosphorylated proteoforms from 14 proteins. Some reported phosphorylated proteoforms are of the same protein and their precursor masses are the same (within an error tolerance). The only difference of these proteoforms is the locations of phosphorylation sites. The 24 phosphorylated proteoforms correspond to 15 distinct precursor masses.

In the first step of the analysis of TopPIC and TopMG, the mouse proteome database was downloaded from the UniProt database (version Nov 13, 2016, 16840 entries) and concatenated with a shuffled decoy database of the same size. We first used TopPIC to search all the deconvoluted spectra against the target-decoy mouse database to identify proteoforms without variable PTMs and unexpected alterations (terminal truncations and N-terminal acetylation are allowed), then used TopMG to search the spectra unidentified by TopPIC against the database to identify phosphorylated proteoforms. In TopPIC, the error tolerances for precursor and fragment masses were set as 10 ppm. In the ASF algorithms, the parameter *h* was set as 1 and the error tolerance for computing filtering scores was set as 10 ppm. In TopMG, the error tolerances for precursor and fragment masses were set as 10 ppm and 0.1 Da respectively, and phosphorylation was used as the variable PTM. Other parameter settings of TopPIC and TopMG are given in Tables S10 and S11 in the supplementary material. With a 5% proteoform-level FDR, TopPIC identified 122 proteoforms from 105 proteins, and TopMG identified 45 proteoforms, including 41 phosphorylated proteoforms from 27 proteins and 4 proteoforms without phosphorylation sites (Tables S12-S14 in the supplementary material). The reason that the 4 unmodified proteoforms were missed by TopPIC is that TopPIC used a more stringent error tolerance for fragment masses compared with TopMG. Most of the identified phosphorylated proteoforms contain ≤ 3 phosphorylation sites (Fig. S19(a) in the supplementary material).

A total of 21 proteoforms without variable PTMs (some may contain terminal truncations and N-terminal acetylation) were identified by both ProSightPC and TopPIC. In addition, TopPIC identified 101 proteoforms missed by ProSightPC (Fig. S20(a) in the supplementary material).

Because the spectral scan numbers of the proteoforms reported by ProSightPC were not available, we matched the molecular masses of the proteoforms to the precursor masses of the spectra reported by MS-Deconv with an error tolerance 2.2 Da to find candidate PrSMs. Of the 20 proteoforms missed by TopPIC, MS-Deconv failed to report corresponding deconvoluted spectra for 4 proteoforms. The molecular masses of the other 16 proteoforms were matched to the precursor masses of 242 deconvoluted spectra, but their corresponding PrSMs were not reported by TopPIC because their *E*-values were not highly significant. One main reason that ProSightPC missed many proteoforms identified by TopPIC is that truncations were not allowed in the first step of the ProSightPC analysis.

ProSightPC reported several proteoforms with the same molecular mass, but different PTM sites. Because it is a challenging problem to confidently localize PTM sites in top-down spectral identification, we decided not to directly compare proteoforms reported by the two tools. If a proteoform reported by ProSightPC and a proteoform reported by TopMG are of the same protein and have the same precursor mass (within an error tolerance), we say the two proteoforms match. We compared the numbers of distinct precursor masses corresponding to the proteoforms, not the numbers of proteoforms, reported by ProSightPC and TopMG. A total of 38 and 15 distinct precursor masses were reported by TopMG and ProSightPC, respectively. Only one phosphorylated proteoform (corresponding to one precursor mass) was reported by both TopMG and ProSightPC (Fig. S21(a)). Of the remaining 23 phosphorylated proteoforms (14 precursor masses) reported by ProSightPC, 4 did not have matched deconvoluted spectra reported by MS-Deconv, and 19 were matched to deconvoluted spectra, but their corresponding PrSMs were not reported by TopMG. ProSightPC missed many proteoforms reported by TopMG because the proteoform database (data warehouse) used in ProSightPC was incomplete. The proteoforms identified by TopMG include 37 highly confident ones with an *E*-value smaller than 10^{-10} (Fig. S19(b) in the supplementary material). Four proteoforms with significant *E*-values are provided in Fig. S22-S25 in the supplementary material. These identified proteoforms show that TopMG is efficient in identifying novel phosphorylated proteoforms.

Human proteoforms In the second step of the ProSightPC analysis, the absolute mass and biomarker modes were used to search the spectra unidentified in the first step against a human proteoform database, which was built based on the human RefSeq database and protein annotations. The error tolerance for precursor masses was set as 2.2 Da in the absolute mass mode and 10 ppm in the biomarker mode; the error tolerance for fragment masses was set as 10 ppm in the two search modes. With a *p*-value cutoff 10^{-10}, ProSightPC identified 685 proteoforms from 150 proteins[1], including 147 proteoforms without PTMs (N-terminal acetylation is allowed) and 98 phosphorylated proteoforms from 26 proteins. The 98 phosphorylated proteoforms are matched to 35 distinct precursor masses.

[1] The supplementary Table S1 in Ref. [43] shows that the second step identified a mouse protein RS30, which may be an error in the table.

In the second step of the analysis of TopPIC and TopMG, the human proteome database (version Jul 9, 2016, 20191 entries) was downloaded from UniProt and concatenated with a shuffled decoy database with the same size. Using the same parameters in the first step, the spectra unidentified in the first step were searched against the human target-decoy database using TopPIC and TopMG. TopPIC identified 265 proteoforms from 190 proteins without variable PTMs, and TopMG identified 91 proteoforms from 64 proteins, including 82 phosphorylated proteoforms from 59 proteins (Tables S15-S17 in the supplementary material). Similar to the first step, most of the identified phosphorylated proteoforms contain ≤ 3 phosphorylation sites (Fig. S26(a) in the supplementary material).

The human database search of TopPIC identified 85 of the 147 human proteoforms without PTMs (except for terminal truncations and N-terminal acetylation) reported by ProSightPC (Fig. S20(b) in the supplementary material). Of the 62 proteoforms missed by TopPIC, 13 were identified by TopPIC in the mouse database search because they are the same as their homologous mouse proteins. Similar to mouse proteoforms, the main reasons for the remaining 49 proteoforms missed by TopPIC are the missing of matched deconvoluted spectra and large *E*-values of PrSMs. TopPIC also identified 180 proteoforms missed by ProSightPC.

A total of 80 and 35 distinct precursor masses were reported by TopMG and ProSightPC, including 14 ones reported by both the two tools (Fig. S21(b)). The proteoforms identified by TopMG include 47 proteoforms with an *E*-value smaller than 10^{-10} (Fig. S26(b) in the supplementary material). Four proteoforms with significant *E*-values are provided in Fig. S27-S30 in the supplementary material. Similar to the comparison on mouse phosphorylated proteoforms, TopMG identified many phosphorylated human proteoforms missed by the absolute mass and biomarker modes of ProSightPC. All the annotated PrSMs reported by TopPIC and TopMG can be found in the supplementary material.

4 Discussion and conclusions

In this paper, we proposed two ASF algorithms for protein filtration in proteoform identification by top-down MS and evaluated the performances of the ASF algorithms as well as two tag-based and two UPF-based filtering algorithms on simulated and real top-down MS data sets. The experimental results showed that the UPF-based filtering algorithms outperformed the tag-based algorithms and that the ASF algorithms achieved the best performance among the 6 evaluated algorithms in filtration efficiency. The ASF algorithms are efficient when the target proteoform contains truncations as well as many variable PTMs and/or unknown alterations. Specifically, the filtration efficiency of ASF-DIAGONAL is much higher than other methods for spectra with low sequence coverage. Although the ASF algorithms are the slowest, their speed is still acceptable in proteoform identification.

Both ASF-RESTRICT and ASF-DIAGONAL use approximate spectra in protein filtration, but they are designed for different scenarios. ASF-RESTRICT has a smaller search space than ASF-DIAGONAL. While the filtration efficiency of ASF-RESTRICT depends on if the corresponding proteoform of the approximate spectrum contains a long unmodified prefix or suffix, the filtration efficiency of ASF-DIAGONAL depends on if the corresponding proteoform of the approximate

spectrum contains a long unmodified fragment (a prefix, a suffix, or an internal one). In practice, we suggest combining the two algorithms to achieve good filtration efficiency.

The parameters *h*, *f*, and *k* determine the search space, running time, and filtration efficiency of the ASF algorithms. When *h*, *f*, and *k* increases, the search space and running time increase. The experimental results demonstrate that using one variable PTM site in approximate spectrum generation (*h* = 1) significantly improves filtration efficiency for complex proteoforms with multiple various PTMs compared with UPF-based methods. While using *h* = 2 achieves marginal improvement in filtration efficiency compared with *h* = 1, it significantly increases the running time. We suggest using *h* = 1 in most cases. When only one or two types of variable PTMs are used (*f* = 1 or 2) and many proteoforms are highly modified, *h* = 2 can be used to further improve filtration efficiency. To guarantee that the ASF algorithms are fast in protein filtration, we suggest that the settings of *k* and *f* should be no more than 5.

The ASF algorithms are proposed for proteoform identification in proteome-level proteomics studies in which all proteoforms in the sample are analyzed in an MS experiment. The types of PTMs of interest are known in many proteome-level proteomics studies. For example, phosphorylation is the PTM of interest and chosen as the variable PTM in the studies of phosphoproteins. In the discovery mode analysis, the types of PTMs of interest are unknown and it is a challenging problem to anticipate the types of PTMs that will be identified in proteoforms. To solve the problem, we first use spectral alignment algorithms, such as TopPIC, to identify proteoforms with mass shifts corresponding to unexpected alterations. If the number of occurrences of a specific mass shift, e.g. 80 Da, in identified proteoforms is large and the mass shift is explained by a PTM (80 Da is explained by phosphorylation), then we use the PTM as a variable one in the second round of database search to find proteoforms with the PTM.

The number of variable PTM types needs to be small to guarantee the fast speed of the ASF algorithms. A proteome level MS analysis may identify more than 10 types of PTMs, but each proteoform often contains only one or two types of PTMs. To identify these proteoforms, we can perform multiple rounds of database searches, and a small number of variable PTM types are selected in each round.

A proteoform may contain various alterations including terminal truncations, sequence mutations, fixed PTMs, variable PTMs, and unexpected alterations. The ASF algorithms are capable of filtering spectra of proteoforms with truncations, fixed PTMs, variable PTMs, and unexpected alterations. When sample specific protein databases are not available, sequence mutations are treated as unexpected alterations in protein filtration. When RNA-Seq data of the sample are available, sequence mutations obtained from RNA-Seq data can be incorporated into sample specific protein databases to improve filtration efficiency. When the target proteoform contains many variable PTM sites, most of them are treated as unexpected alterations in filtration because approximate spectra usually remove only one or two variable PTM sites (*h* = 1 or 2) in the proteoform.

Unexpected alterations and the alterations that are treated as unexpected ones in filtration are called filtration blind alterations. The number and locations of filtration blind alterations affect

the filtration efficiency of the ASF algorithms. In general, the filtration efficiency decreases when the number of filtration blind alterations increases. ASF-DIAGONAL filters proteins using a long unmodified protein fragment. When a proteoform with many filtration bind alterations has a long fragment free of filtration bind alterations, it is highly possible that ASF-DIAGONAL is efficient for the proteoform. Similarly, when a proteoform with many filtration blind alterations contains a long prefix or suffix free of filtration blind alterations, it is highly possible that ASF-RESTRICT is efficient for the proteoform.

In proteome-level proteomics studies, proteoforms can be divided into three groups: (1) proteoforms with only variable PTMs, (2) proteoforms with only filtration blind alterations, and (3) proteoforms with both variable PTMs and filtration blind alterations. The ASF algorithms are designed to improve the sensitivity in proteoform identification in groups (1) and (3), but not in group (2). That is, the ASF algorithms work well for proteoforms with only variable PTMs, and those with both variable PTMs and unexpected alterations, not for proteoforms with only unexpected alterations.

In the ASF algorithms, the query spectrum is transformed into an approximate spectrum to reduce the number of variable PTMs in the match between the target database sequence and the spectrum. An alternative method is to incorporate variable PTMs into database sequences to generate a proteoform database. This approach has been widely used in PTM identification in bottom-up MS, but it is inefficient in top-down MS. Proteoforms analyzed in top-down MS are generally longer than peptides in bottom-up MS. Because long proteins often contain many possible modification sites, the size of a proteoform database may be extremely large. For example, when phosphorylation is the only variable PTM and one or two PTM sites (h = 2) are incorporated into each proteoform, the size of the proteoform database increases by more than 100 times compared with the original one.

The proposed ASF algorithms have some limitations. The first limitation is that the running time of the algorithms is an exponential function of the parameter h. In practice, a small number h (h = 1 or 2) is used to reduce the running time of the algorithms, limiting its ability to identify complex proteoforms with many variable PTM sites. The second limitation is that the ASF algorithms are inefficient for proteoforms with many PTM types. Using a large number (> 5) of variable PTM types significantly increases the running time of the algorithms. The third limitation is that peak intensities are ignored in computing diagonal scores and restricted diagonal scores. Incorporating peak intensities into similarity scores can further improve the performance of the filtering algorithms.

Availability The proposed ASF algorithms have been integrated into TopMG, which is available at http://proteomics.informatics.iupui.edu/software/topmg/. The source code is available at https://github.com/toppic-suite/toppic-suite.

Acknowledgement

The research was supported by the National Institute of General Medical Sciences, National Institutes of Health (NIH) through Grant R01GM118470. The authors declare no competing financial interest.

Reference

[1] L. M. Smith, N. L. Kelleher, C. f. T. D. Proteomics, Proteoform: a single term describing protein complexity. *Nature Methods* 2013, *10*, 186-187.
[2] X. Dong, C. A. Sumandea, Y.-C. Chen, M. L. Garcia-Cazarin, J. Zhang, C. W. Balke, M. P. Sumandea, Y. Ge, Augmented phosphorylation of cardiac troponin I in hypertensive heart failure. *Journal of Biological Chemistry* 2012, *287*, 848-857.
[3] S. Peleg, F. Sananbenesi, A. Zovoilis, S. Burkhardt, S. Bahari-Javan, R. C. Agis-Balboa, P. Cota, J. L. Wittnam, A. Gogol-Doering, L. Opitz, G. Salinas-Riester, M. Dettenhofer, H. Kang, L. Farinelli, W. Chen, A. e. Fischer, ,, Altered histone acetylation is associated with age-dependent memory impairment in mice. *Science* 2010, *328*, 753-756.
[4] B. A. Garcia, J. J. Pesavento, C. A. Mizzen, N. L. Kelleher, Pervasive combinatorial modification of histone H3 in human cells. *Nature methods* 2007, *4*, 487-489.
[5] N. L. Young, P. A. DiMaggio, M. D. Plazas-Mayorca, R. C. Baliban, C. A. Floudas, B. A. Garcia, High throughput characterization of combinatorial histone codes. *Molecular & Cellular Proteomics* 2009, *8*, 2266-2284.
[6] A. D. Catherman, O. S. Skinner, N. L. Kelleher, Top Down proteomics: facts and perspectives. *Biochemical and Biophysical Research Communications* 2014, *445*, 683-693.
[7] L. Zamdborg, R. D. LeDuc, K. J. Glowacz, Y.-B. Kim, V. Viswanathan, I. T. Spaulding, B. P. Early, E. J. Bluhm, S. Babai, N. L. Kelleher, ProSight PTM 2.0: improved protein identification and characterization for top down mass spectrometry. *Nucleic Acids Research* 2007, *35*, W701-W706.
[8] A. M. Frank, J. J. Pesavento, C. A. Mizzen, N. L. Kelleher, P. A. Pevzner, Interpreting top-down mass spectra using spectral alignment. *Analytical Chemistry* 2008, *80*, 2499-2505.
[9] Y. S. Tsai, A. Scherl, J. L. Shaw, C. L. MacKay, S. A. Shaffer, P. R. R. Langridge-Smith, D. R. Goodlett, Precursor ion independent algorithm for top-down shotgun proteomics. *Journal of the American Society for Mass Spectrometry* 2009, *20*, 2154-2166.
[10] N. M. Karabacak, L. Li, A. Tiwari, L. J. Hayward, P. Hong, M. L. Easterling, J. N. Agar, Sensitive and specific identification of wild type and variant proteins from 8 to 669 kDa using top-down mass spectrometry. *Molecular & Cellular Proteomics* 2009, *8*, 846-856.
[11] W. Tong, R. Theberge, G. Infusini, D. H. Perlman, C. E. Costello, M. E. McComb, *Proceedings of the 57th American Society Conference on Mass Spectrometry and Allied Topics, Philadelphia, PA* 2009.
[12] X. Liu, Y. Sirotkin, Y. Shen, G. Anderson, Y. S. Tsai, Y. S. Ting, D. R. Goodlett, R. D. Smith, V. Bafna, P. A. Pevzner, Protein identification using top-down spectra. *Molecular & Cellular Proteomics* 2012, *11*, M111.008524.
[13] M. Bern, Y. J. Kil, C. Becker, Byonic: advanced peptide and protein identification software. *Current Protocols in Bioinformatics* 2012, *Chapter 13*, Unit 13.20.
[14] L. Li, Z. Tian, Interpreting raw biological mass spectra using isotopic mass-to-charge ratio and envelope fingerprinting. *Rapid Communications in Mass Spectrometry* 2013, *27*, 1267-1277.
[15] X. Liu, S. Hengel, S. Wu, N. Tolic, L. Pasa-Tolic, P. A. Pevzner, Identification of ultramodified proteins using top-down tandem mass spectra. *Journal of Proteome Research* 2013, *12*, 5830-5838.
[16] Q. Kou, L. Xun, X. Liu, TopPIC: a software tool for top-down mass spectrometry-based proteoform identification and characterization. *Bioinformatics* 2016, *32*, 3495-3497.
[17] R. X. Sun, L. Luo, L. Wu, R. M. Wang, W. F. Zeng, H. Chi, C. Liu, S. M. He, pTop 1.0: A High-Accuracy and High-Efficiency Search Engine for Intact Protein Identification. *Analytical Chemistry* 2016, *88*, 3082-3090.
[18] Q. Kou, S. Wu, N. Tolić, L. Paša-Tolić, Y. Liu, X. Liu, A mass graph-based approach for the identification of modified proteoforms using top-down tandem mass spectra. *Bioinformatics* 2016, *33*, 1309-1316.

[19] W. Cai, H. Guner, Z. R. Gregorich, A. J. Chen, S. Ayaz-Guner, Y. Peng, S. G. Valeja, X. Liu, Y. Ge, MASH Suite Pro: A comprehensive software tool for top-down proteomics. *Molecular & Cellular Proteomics* 2016, *15*, 703-714.
[20] M. R. Shortreed, B. L. Frey, M. Scalf, R. A. Knoener, A. J. Cesnik, L. M. Smith, Elucidating Proteoform Families from Proteoform Intact-Mass and Lysine-Count Measurements. *Journal of Proteome Research* 2016, *15*, 1213-1221.
[21] J. Park, P. D. Piehowski, C. Wilkins, M. Zhou, J. Mendoza, G. M. Fujimoto, B. C. Gibbons, J. B. Shaw, Y. Shen, A. K. Shukla, R. J. Moore, T. Liu, V. A. Petyuk, N. Tolić, Pa\vs, a-Toli\'c,,, Ljiljana, R. D. Smith, S. H. Payne, S. Kim, Informed-Proteomics: open-source software package for top-down proteomics. *Nature Methods* 2017, *14*, 909-914.
[22] J. M. Chick, D. Kolippakkam, D. P. Nusinow, B. Zhai, R. Rad, E. L. Huttlin, S. P. Gygi, A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nature Biotechnology* 2015, *33*, 743-749.
[23] M. Mann, M. Wilm, Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Analytical chemistry* 1994, *66*, 4390-4399.
[24] S. Tanner, H. Shu, A. Frank, L. C. Wang, E. Zandi, M. Mumby, P. A. Pevzner, V. Bafna, InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Analytical Chemistry* 2005, *77*, 4626-4639.
[25] A. Frank, S. Tanner, V. Bafna, P. Pevzner, Peptide sequence tags for fast database search in mass-spectrometry. *Journal of Proteome Research* 2005, *4*, 1287-1295.
[26] X. Cao, A. I. Nesvizhskii, Improved sequence tag generation method for peptide identification in tandem mass spectrometry. *Journal of Proteome Research* 2008, *7*, 4422-4434.
[27] D. L. Tabb, Z.-Q. Ma, D. B. Martin, A.-J. L. Ham, M. C. Chambers, DirecTag: accurate sequence tags from peptide MS/MS through statistical scoring. *Journal of Proteome Research* 2008, *7*, 3838-3846.
[28] S. Kim, N. Gupta, N. Bandeira, P. A. Pevzner, Spectral dictionaries integrating de novo peptide sequencing with database search of tandem mass spectra. *Molecular & Cellular Proteomics* 2009, *8*, 53-69.
[29] K. Jeong, S. Kim, N. Bandeira, P. A. Pevzner, Gapped spectral dictionaries and their applications for database searches of tandem mass spectra. *Molecular & Cellular Proteomics* 2011, *10*, M110-002220.
[30] F. Deng, L. Wang, X. Liu, An efficient algorithm for the blocked pattern matching problem. *Bioinformatics* 2014, *31*, 532-538.
[31] Y. Shen, N. Tolic, K. K. Hixson, S. O. Purvine, G. A. Anderson, R. D. Smith, De novo sequencing of unique sequence tags for discovery of post-translational modifications of proteins. *Analytical Chemistry* 2008, *80*, 7742-7754.
[32] X. Liu, A. Mammana, V. Bafna, Speeding up tandem mass spectral identification using indexes. *Bioinformatics* 2012, *28*, 1692-1697.
[33] H. Chi, K. He, B. Yang, Z. Chen, R.-X. Sun, S.-B. Fan, K. Zhang, C. Liu, Z.-F. Yuan, Q.-H. Wang, S.-Q. Liu, M.-Q. Dong, S.-M. He, pFind-Alioth: A novel unrestricted database search algorithm to improve the interpretation of high-resolution MS/MS data. *Journal of Proteomics* 2015, *125*, 89-97.
[34] A. T. Kong, F. V. Leprevost, D. M. Avtonomov, D. Mellacheruvu, A. I. Nesvizhskii, MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nature Methods* 2017, *14*, 513-520.
[35] D. M. Horn, R. A. Zubarev, F. W. McLafferty, Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *Journal of the American Society for Mass Spectrometry* 2000, *11*, 320-332.
[36] V. Zabrouskov, M. W. Senko, Y. Du, R. D. Leduc, N. L. Kelleher, New and automated MSn approaches for top-down identification of modified proteins. *Journal of the American Society for Mass Spectrometry* 2005, *16*, 2027-2038.

[37] A. M. Mayampurath, N. Jaitly, S. O. Purvine, M. E. Monroe, K. J. Auberry, J. N. Adkins, R. D. Smith, DeconMSn: a software tool for accurate parent ion monoisotopic mass determination for tandem mass spectra. *Bioinformatics* 2008, *24*, 1021-1023.
[38] P. C. Carvalho, T. Xu, X. Han, D. Cociorva, V. C. Barbosa, J. R. Yates III, YADA: a tool for taking the most out of high-resolution spectra. *Bioinformatics* 2009, *25*, 2734-2736.
[39] X. Liu, Y. Inbar, P. C. Dorrestein, C. Wynne, N. Edwards, P. Souda, J. P. Whitelegge, V. Bafna, P. A. Pevzner, Deconvolution and database search of complex tandem mass spectra of intact proteins: a combinatorial approach. *Molecular & Cellular Proteomics* 2010, *9*, 2772-2782.
[40] G. W. Slysz, E. S. Baker, A. R. Shah, N. Jaitly, G. A. Anderson, R. D. Smith, *Proceedings of the 58th American Society Conference on Mass Spectrometry and Allied Topics* 2010.
[41] Q. Kou, S. Wu, X. Liu, A new scoring function for top-down spectral deconvolution. *BMC Genomics* 2014, *15*, 1140.
[42] Z. Tian, N. Tolic, R. Zhao, R. J. Moore, S. M. Hengel, E. W. Robinson, D. L. Stenoien, S. Wu, R. D. Smith, L. Pasa-Tolic, Enhanced top-down characterization of histone post-translational modifications. *Genome Biology* 2012, *13*, R86.
[43] I. Ntai, R. D. LeDuc, R. T. Fellers, P. Erdmann-Gilmore, S. R. Davies, J. Rumsey, B. P. Early, P. M. Thomas, S. Li, P. D. Compton, M. J. C. Ellis, K. V. Ruggles, Feny\"o, , David, E. S. Boja, H. Rodriguez, R. R. Townsend, N. L. Kelleher, Integrated bottom-up and top-down proteomics of patient-derived breast tumor xenografts. *Molecular & Cellular Proteomics* 2016, *15*, 45-56.
[44] P. Mertins, F. Yang, T. Liu, D. R. Mani, V. A. Petyuk, M. A. Gillette, K. R. Clauser, J. W. Qiao, M. A. Gritsenko, R. J. Moore, D. A. Levine, R. Townsend, P. Erdmann-Gilmore, J. E. Snider, S. R. Davies, K. V. Ruggles, D. Fenyo, R. T. Kitchens, S. Li, N. Olvera, F. Dao, H. Rodriguez, D. W. Chan, D. Liebler, F. White, K. D. Rodland, G. B. Mills, R. D. Smith, A. G. Paulovich, M. Ellis, S. A. Carr, Ischemia in tumors induces early and sustained phosphorylation changes in stress kinase pathways but does not affect global protein levels. *Molecular & Cellular Proteomics* 2014, *13*, 1690-1704.
[45] L. Ding, M. J. Ellis, S. Li, D. E. Larson, K. Chen, J. W. Wallis, C. C. Harris, M. D. McLellan, R. S. Fulton, L. L. Fulton, R. M. Abbott, J. Hoog, D. J. Dooling, D. C. Koboldt, H. Schmidt, J. Kalicki, Q. Zhang, L. Chen, L. Lin, M. C. Wendl, J. F. McMichael, V. J. Magrini, L. Cook, S. D. McGrath, T. L. Vickery, E. Appelbaum, K. DeSchryver, S. Davies, T. Guintoli, L. Lin, R. Crowder, Y. Tao, J. E. Snider, S. M. Smith, A. F. Dukes, G. E. Sanderson, C. S. Pohl, K. D. Delehaunty, C. C. Fronick, K. A. Pape, J. S. Reed, J. S. Robinson, J. S. Hodges, W. Schierding, N. D. Dees, D. Shen, D. P. Locke, M. E. Wiechert, J. M. Eldred, J. B. Peck, B. J. Oberkfell, J. T. Lolofie, F. Du, A. E. Hawkins, M. D. O'Laughlin, K. E. Bernard, M. Cunningham, G. Elliott, M. D. Mason, D. M. T. Jr, J. L. Ivanovich, P. J. Goodfellow, C. M. Perou, G. M. Weinstock, R. Aft, M. Watson, T. J. Ley, R. K. Wilson, E. R. Mardis, Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* 2010, *464*, 999-1005.
[46] S. Li, D. Shen, J. Shao, R. Crowder, W. Liu, A. Prat, X. He, S. Liu, J. Hoog, C. Lu, L. Ding, O. L. Griffith, C. Miller, D. Larson, R. S. Fulton, M. Harrison, T. Mooney, J. F. McMichael, J. Luo, Y. Tao, R. Goncalves, C. Schlosberg, J. F. Hiken, L. Saied, C. Sanchez, T. Giuntoli, C. Bumb, C. Cooper, R. T. Kitchens, A. Lin, C. Phommaly, S. R. Davies, J. Zhang, M. S. Kavuri, D. McEachern, Y. Y. Dong, C. Ma, T. Pluard, M. Naughton, R. Bose, R. Suresh, R. McDowell, L. Michel, R. Aft, W. Gillanders, K. DeSchryver, R. K. Wilson, S. Wang, G. B. Mills, A. Gonzalez-Angulo, J. R. Edwards, C. Maher, C. M. Perou, E. R. Mardis, M. J. Ellis, Endocrine-therapy-resistant ESR1 variants revealed by genomic characterization of breast-cancer-derived xenografts. *Cell Reports* 2013, *4*, 1116-1130.
[47] T. U. Consortium, UniProt: a hub for protein information. *Nucleic Acids Research* 2015, *43*, D204-D212.
[48] S. Kim, N. Gupta, P. A. Pevzner, Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *Journal of Proteome Research* 2008, *7*, 3354-3363.
[49] X. Liu, M. W. Segar, S. C. Li, S. Kim, Spectral probabilities of top-down tandem mass spectra. *BMC genomics* 2014, *15*, S9.

Tables

Table 1 Comparison of the 6 filtering algorithms in the filtration efficiency rate using the 3205 histone H3 PrSMs and the 1087 histone H4 PrSMs

	H3			H4		
	# efficiently filtered PrSMs	Efficiency rate	Time (minutes)	# efficiently filtered PrSMs	Efficiency rate	Time (minutes)
TAG-LONG	210	6.6%	91.8	563	51.8%	73.9
TAG-VAR	415	13.0%	92.1	583	53.6%	73.9
UPF-RESTRICT	2019	63.0%	35.7	1052	96.8%	11.2
UPF-DIAGONAL	940	29.3%	507.4	1014	93.3%	87.7
ASF-RESTRICT	2313	72.2%	400.7	1080	99.3%	150.8
ASF-DIAGONAL	1235	38.5%	4642.0	1036	95.3%	1307.4

Figures

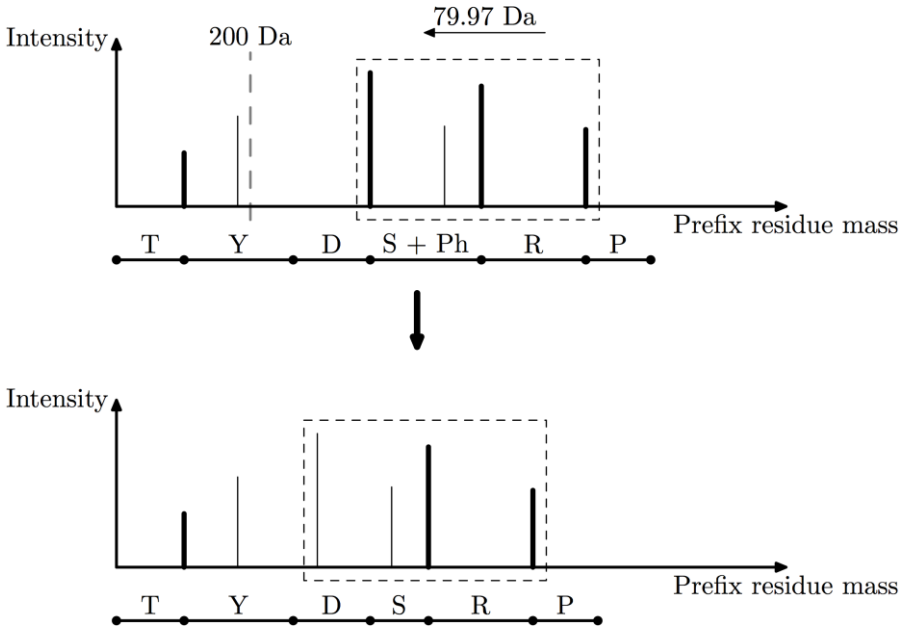

Figure 1: A prefix residue mass spectrum (top) of the proteoform TYDS[Ph]RP with a phosphorylation site on the serine residue is transformed into an approximate prefix residue mass spectrum (bottom) of the unmodified protein TYDSPR. In the top spectrum, each peak represents a possible prefix residue mass extracted from the experimental spectrum, and bold peaks are those mapped to theoretical prefix residue masses of the proteoform TYDS[Ph]RP. The prefix residue mass 200 Da is a guessed prefix residue mass for the modification site. All peaks (in the box) with a mass larger than 200 Da are shifted to the left by 79.97 Da, which is the mass shift of a phosphorylation site. In the bottom spectrum, the two shifted bold peaks in the box are matched to prefix residue masses of TYDSRP, and the left most peak in the box is not matched to any prefix residue mass of TYDSRP because of the error in the estimated 200 Da for the modification site.

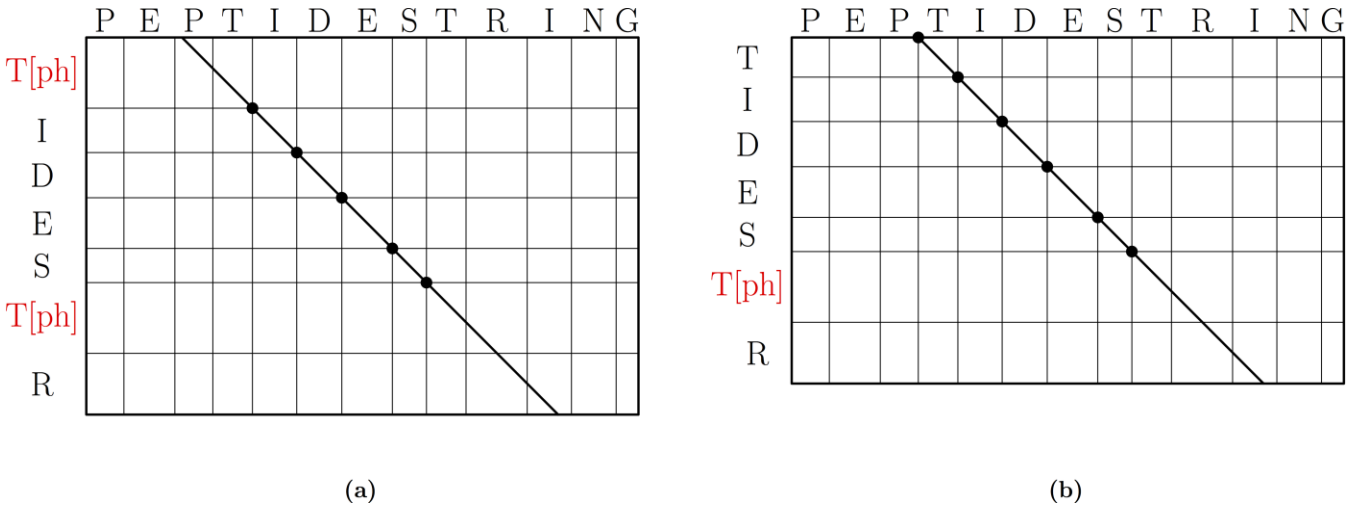

Figure 2: Diagonal scores and restricted diagonal scores. (a) The diagonal score between the prefix residue masses of PEPTIDESTRING and T[Ph]IDEST[Ph]R is 5, corresponding to the 5 dots in the diagonal. The score is obtained by shifting the prefix residue masses of PEPTIDESTRING by -243.18 Da, which equals -mass(PEPT)+mass(T[Ph]). (b) The restricted diagonal score between the prefix residue mass of PEPTIDESTRING and TIDEST[Ph]R is 6. The score is obtained by shifting the prefix residue masses of PEPTIDESTRING by -323.15 Da = -mass(PEP).

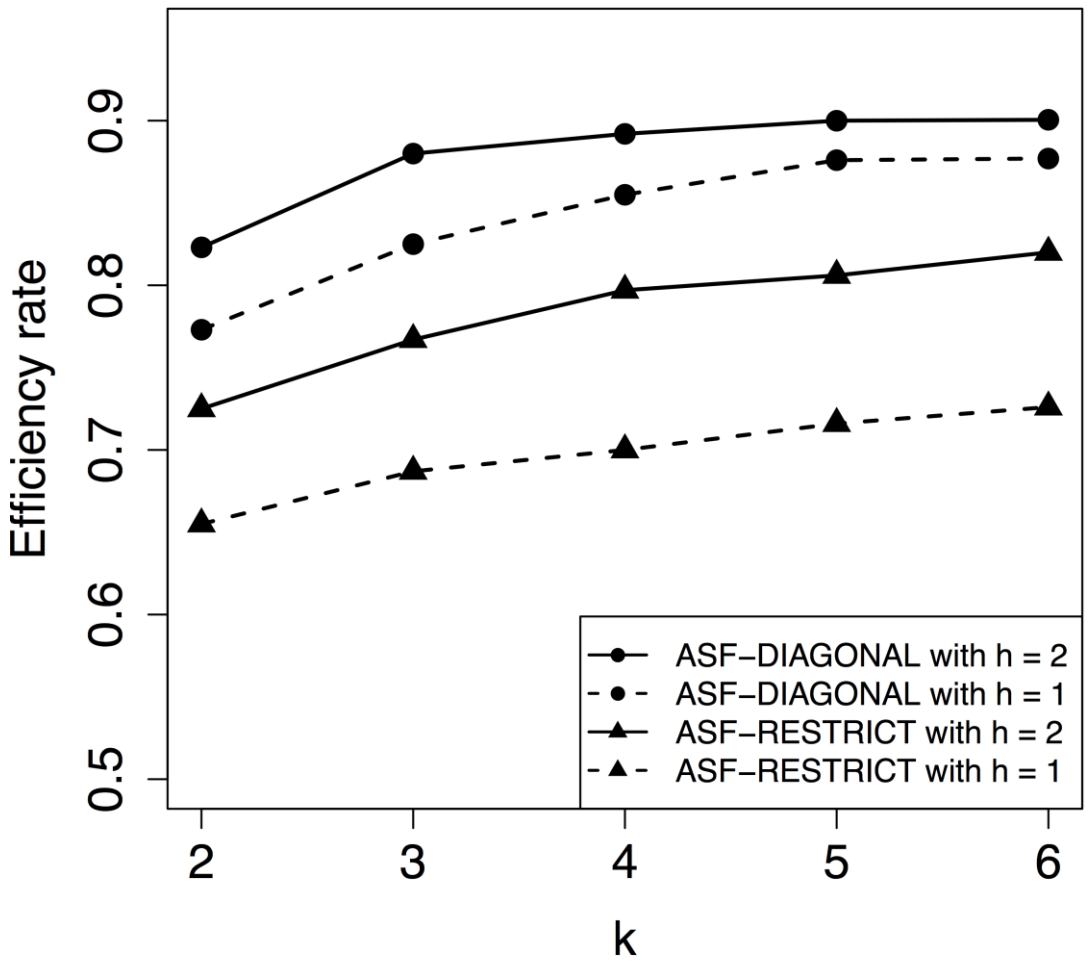

Figure 3: The efficiency rates of the ASF algorithms with various settings k = 2, 3, 4, 5, 6 and h = 1, 2 on the simulated PrSMs with 5 PTMs.

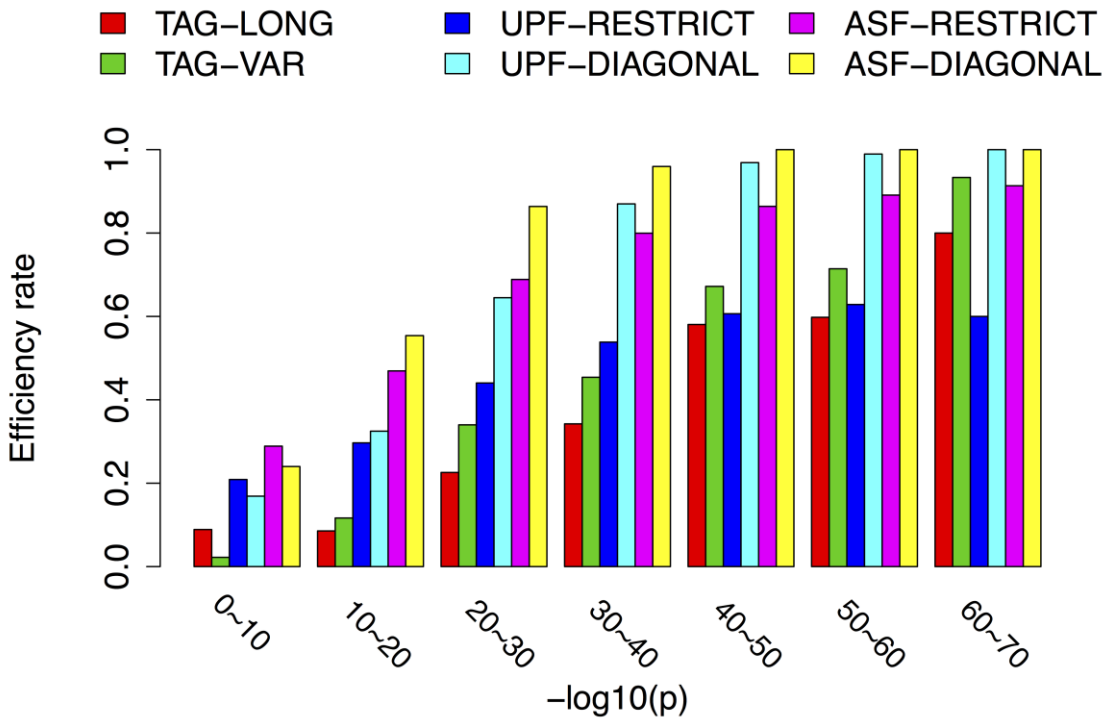

Figure 4: Comparison of the filtration efficiency rates of the TAG-LONG, TAG-VAR, UPF-RESTRICT, UPF-DIAGONAL, ASF-RESTRICT and ASF-DIAGONAL algorithms on the simulated test PrSMs with 5 PTMs. The PrSMs are divided into 7 groups based on their conditional spectral probabilities p, and the efficiency rates for each group are compared.