COMPLEX PROTEOFORM IDENTIFICATION USING TOP-DOWN MASS SPECTROMETRY

Qiang Kou

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the School of Informatics and Computing,
Indiana University
December 2018

Accepted by the Graduate Faculty of Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

Huanmei Wu, PhD, Chair

Xiaowen Liu, PhD

Yunlong Liu, PhD

August 22, 2018

Mohammad Al Hasan, PhD

ACKNOWLEDGMENTS

Without the guidance of my committee and support from my family, I would not survive the long march of the Ph.D. study. There are many people I need to thank for making my five years of study meaningful and enjoyable.

I want to express my most profound gratitude to my advisor, Prof. Xiaowen Liu, for his continuous support, excellent guidance, and immense knowledge through my studies and research. This dissertation can never be finished without his advice and support.

Besides my advisor, I would like to thank the rest of my committee: Prof. Yunlong Liu, Prof. Huanmei Wu, and Prof. Mohammad Al Hasan, for their encouragement and insightful comments.

I also want to thank our collaborators: Prof. Si Wu (the University of Oklahoma), Prof. Liangliang Sun (Michigan State University), Prof. Binhai Zhu (Montana State University), Dr. Nikola Tolić and Dr. Ljiljana Paša-Tolić (Pacific Northwest National Laboratory). Their expertise broadened my view of science.

Last but not the least, I want to thank my family for supporting me during the years of life.

Chapter 3 is in part adapted from *Qiang Kou, Si Wu, and Xiaowen Liu. Systematic evaluation of protein sequence filtering algorithms for proteoform identification using top-down mass spectrometry. Proteomics, 18(3-4):1700306, 2018.* The dissertation author was the primary author responsible for the research of the paper.

Chapter 4 is in part adapted from *Qiang Kou, Si Wu, Nikola Tolić, Ljiljana Paša-Tolić, Yunlong Liu, and Xiaowen Liu. A mass graph-based approach for the identification of modified proteoforms using top-down tandem mass spectra. Bioinformatics, 33(9), pp.1309-1316, 2017.* The dissertation author was the primary author responsible for the research of the paper.

Chapter 5 is in preparation for publication as *Qiang Kou, Si Wu, Liangliang Sun, and Xiaowen Liu. A Markov chain Monte Carlo method for estimating the statistical significance of top-down mass spectrometry-based proteoform identifications.* The dissertation author was the primary author responsible for the research of the paper.

Qiang Kou

COMPLEX PROTEOFORM IDENTIFICATION USING TOP-DOWN MASS SPECTROMETRY

Proteoforms are distinct protein molecule forms created by variations in genes, gene expression, and other biological processes. Many proteoforms contain multiple primary structural alterations, including amino acid substitutions, terminal truncations, and post-translational modifications. These primary structural alterations play a crucial role in determining protein functions: proteoforms from the same protein with different alterations may exhibit different functional behaviors. Because top-down mass spectrometry directly analyzes intact proteoforms and provides complete sequence information of proteoforms, it has become the method of choice for the identification of complex proteoforms. Although instruments and experimental protocols for top-down mass spectrometry have been advancing rapidly in the past several years, many computational problems in this area remain unsolved, and the development of software tools for analyzing such data is still at its very early stage. In this dissertation, we propose several novel algorithms for challenging computational problems in proteoform identification by top-down mass spectrometry. First, we present two approximate spectrum-based protein sequence filtering algorithms that quickly find a small number of candidate proteins from a large proteome database for a query mass spectrum. Second, we describe mass graph-based alignment algorithms that efficiently identify proteoforms with variable post-translational modifications and/or terminal truncations. Third, we propose a Markov chain Monte Carlo method for estimating the statistical significance of identified proteoform spectrum matches. They are the first efficient algorithms that take into account three types of alterations: variable post-translational modifications, unexpected alterations, and terminal truncations in proteoform identification. As a result, they are more sensitive and powerful than other existing methods that consider only one or two of the three types of alterations. All the proposed algorithms have been incorporated into TopMG, a complete software pipeline for complex proteoform identification. Experimental results showed that TopMG significantly increases the number of identifications than other existing methods in proteome-level top-down mass spectrometry studies.

TopMG will facilitate the applications of top-down mass spectrometry in many areas, such as the identification and quantification of clinically relevant proteoforms and the discovery of new proteoform biomarkers.

Huanmei Wu, PhD, Chair

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ASF	approximate spectrum-based filtering
CID	collision-induced dissociation
CZE	capillary zone electrophoresis
DPR	direct probability redistribution
EC	*Escherichia coli*
ETD	electron-transfer dissociation
ESI	electrospray ionization
FDR	false discovery rate
HCD	higher-energy collisional dissociation
HPLC	high-performance liquid chromatography
LC	liquid chromatography
MCMC	Markov chain Monte Carlo
MS	mass spectrometry
MS/MS	tandem mass spectrometry
ppm	parts per million
PrSM	proteoform spectrum match
PSA	primary structural alteration
PTM	post-translational modification
UPF	unmodified protein fragments

CHAPTER 1

INTRODUCTION

1.1 Proteoforms

Proteoforms [2] are distinct protein forms created by variations in genes, gene expression, and other biological processes. These proteoforms often contain multiple primary structural alterations (PSAs), including amino acid sequence variations, terminal truncations, and post-translational modifications (PTMs). Proteoforms are functional macromolecules in cellular processes, and proteoform functions are primarily determined by PSAs. For example, the combinatorial patterns of PSAs in histone proteins determine their gene regulatory functions [3, 4]. Identification and characterization of these proteoforms aid researchers in answering many questions in basic and translational research [5, 6].

Most protein sequence databases, such as UniProt [7], provide only one reference protein sequence for each gene or transcript isoform even though many proteoforms can be generated from one gene or transcript. Compared with its corresponding reference sequence in the database, a complex proteoform often contains various PSAs (Figure 1.1), which can be divided into five categories: (a) fixed PTMs, which modify every instance of specific residues in the protein sequence, such as carbamidomethylation and carboxymethylation; (b) sequence variations, such as mutations, insertions, and deletions; (c) variable PTMs, which may or may not modify specific residues in the protein sequence, such as phosphorylation and oxidation; (d) terminal truncations, which remove a prefix and/or a suffix of the protein sequence; and (e) unknown mass shifts introduced by unknown PSAs. In Figure 1.1, the proteoform contains four kinds of PSAs: carbamidomethylation is a fixed PTM that modifies every cysteine (C) residue; phosphorylation is a variable PTM that may modify serine (S), threonine (T), and tyrosine (Y) residues, but only one serine (S) residue is modified in the proteoform; the N-terminal truncation removes the prefix "MTE"; multiple sequence variations are also observed, including an amino acid mutation from "R" to "K", an insertion of "KK", a deletion of "G". The differences between the target proteoform and its reference sequence make proteoform identification a challenging computational problem.

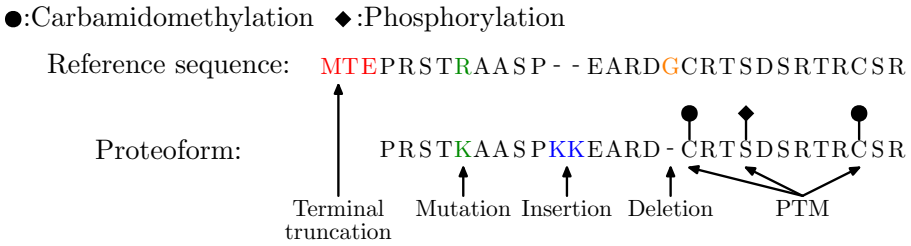

Figure 1.1: Comparison of a complex proteoform and its corresponding reference protein sequence in the database. The proteoform has an N-terminal truncation "MTE", an amino acid mutation from "R" to "K", an insertion of "KK", a deletion of "G", one phosphorylated serine residue, and two modified cysteine residues with carbamidomethylation. Revised from Kou et al. [1].

1.2 Mass spectrometry

Mass spectrometry (MS) is an analytic technique for measuring the mass-to-charge ratios (m/z) of charged particles. It is the *de facto* standard method for high-throughput proteomics studies. Figure 1.2 illustrates the ionization and measurement steps in an MS experiment, in which the ion abundance and m/z value of an ionized protein are measured. The application of MS in biomedical research made rapid progress with the development of the electrospray ionization (ESI) technique [8]. Because ESI generates ions directly from the solution, liquid chromatography (LC) are often used for protein separation and coupled to a mass spectrometer in proteome-level proteomics studies.

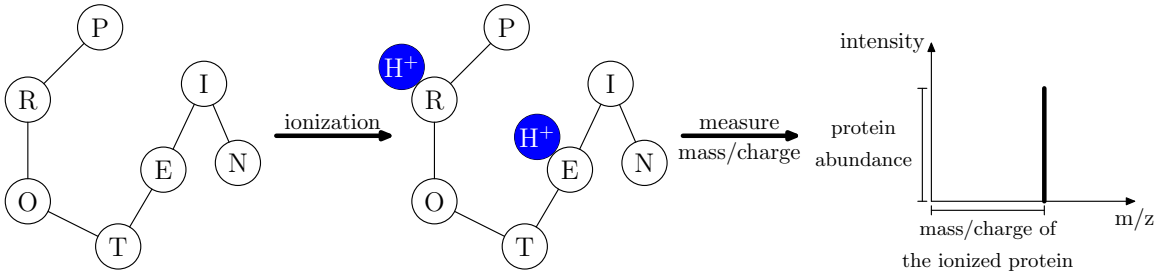

Figure 1.2: An illustration of the ionization and measurement steps in MS. The abundance and mass-to-charge ratio of a protein are measured.

Using only the molecular masses of proteoforms cannot provide sufficient information to infer their corresponding amino acid sequences. The reason is that a molecular mass has many candidate proteoforms that have different amino acid sequences. Tandem mass spectrometry (MS/MS) [9] was introduced to solve this problem by breaking down precursor

ions and measuring m/z values of their fragments ions (Figure 1.3). Two mass analyzers are used in the MS/MS method: the first mass analyzer (MS1) isolates the precursor ions in a fixed range of m/z and store them in a chamber (e.g., an ion trap) where the precursor ions are fragmented. Commonly used fragmentation methods include collision with neutral gas molecules [10, 11] and transferring electrons to positively charged molecules [12]. The resulting fragment ions are analyzed by the second mass analyzer (MS2) to generate an MS/MS spectrum. A precursor ion is often broken into two fragment ions in the process. Because many precursor ions of the same proteoform are collected and the breakage points of precursor ions are not fixed, a list of fragment ions with various breakage points are generated. Ideally, the MS/MS spectrum contains peaks of fragment ions supporting all breakage points of the proteoform, providing enough information for proteoform identification and characterization. The MS/MS method achieved great success in proteomics studies [8, 13, 14] because of its accuracy and high-throughput. The fast development of high-resolution mass spectrometers makes in-depth profiling of complex proteomes not only feasible [15] but also time-efficient [16].

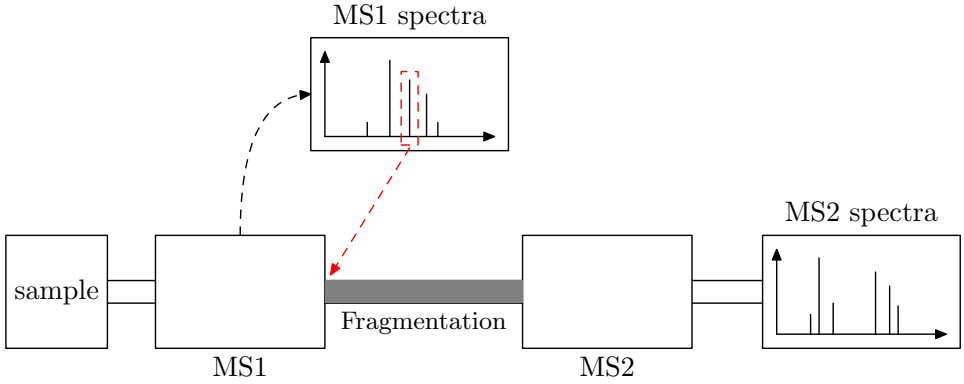

Figure 1.3: An illustration of tandem mass spectrometry (MS/MS) in which two mass analyzers are used. The first mass analyzer (MS1) isolates the target precursor ions (in the red dotted box) from other ions based on m/z values. Then the selected precursor ions are fragmented and analyzed by the second mass analyzer (MS2).

1.3 MS-based proteoform identification

MS methods in proteomics research can be roughly divided into two categories: bottom-up MS and top-down MS. In bottom-up MS, proteins are digested with a protease, such as

trypsin, before MS analysis. The digestion step results in a mixture of short peptides. On the other hand, top-down MS skips the digestion step and directly analyzes intact proteoforms [17]. This gives top-down MS unique advantages in identifying complex proteoforms with multiple PSAs. Fragment ion series in top-down MS/MS spectra provide essential information for identifying and localizing PSAs.

A top-down MS/MS spectrum contains a list of peaks (Figure 1.4(a)), 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 proteoform being studied. The first step in top-down spectral interpretation is usually spectral deconvolution [18–25], which converts fragment ion peaks of various charge states and isotopomers into neutral monoisotopic fragment masses (Figure 1.4).

Let DB be a protein sequence database and Ω a set of variable PTMs (sequence variations can be handled as variable PTMs). The set of all possible proteoforms generated from sequences in DB with variable PTMs in Ω and/or terminal truncations is denoted by $DB(\Omega)$. Given a deconvoluted MS/MS spectrum S and a sequence database DB, the proteoform identification problem is to find the proteoform $F \in DB(\Omega)$, which can best explain S. Various scoring functions [26] for peptide spectrum matches in bottom-up MS can be applied to measure the similarity of the proteoform spectrum matches (PrSM) (F, S). In this thesis, we evaluate (F, S) using the *shared mass counting score* which counts the number of neutral masses in S explained by the theoretical neutral fragment masses of F.

Database search is the dominant method for this problem, where top-down MS/MS spectra are searched against a protein sequence database or an annotated database for spectral identification [1, 18, 27–39]. A list of available software tools for proteoform identification using top-down MS is shown in Table 1.1.

Extended proteoform databases and spectral alignment are the two main strategies in database search. ProSightPC [27] constructs a "shotgun annotated" proteoform database containing known modified proteoforms, and efficiently identifies proteoforms in the database. Because it only includes the commonly observed proteoforms to keep the database size manageable, its ability to identify uncommon or novel proteoforms is limited.

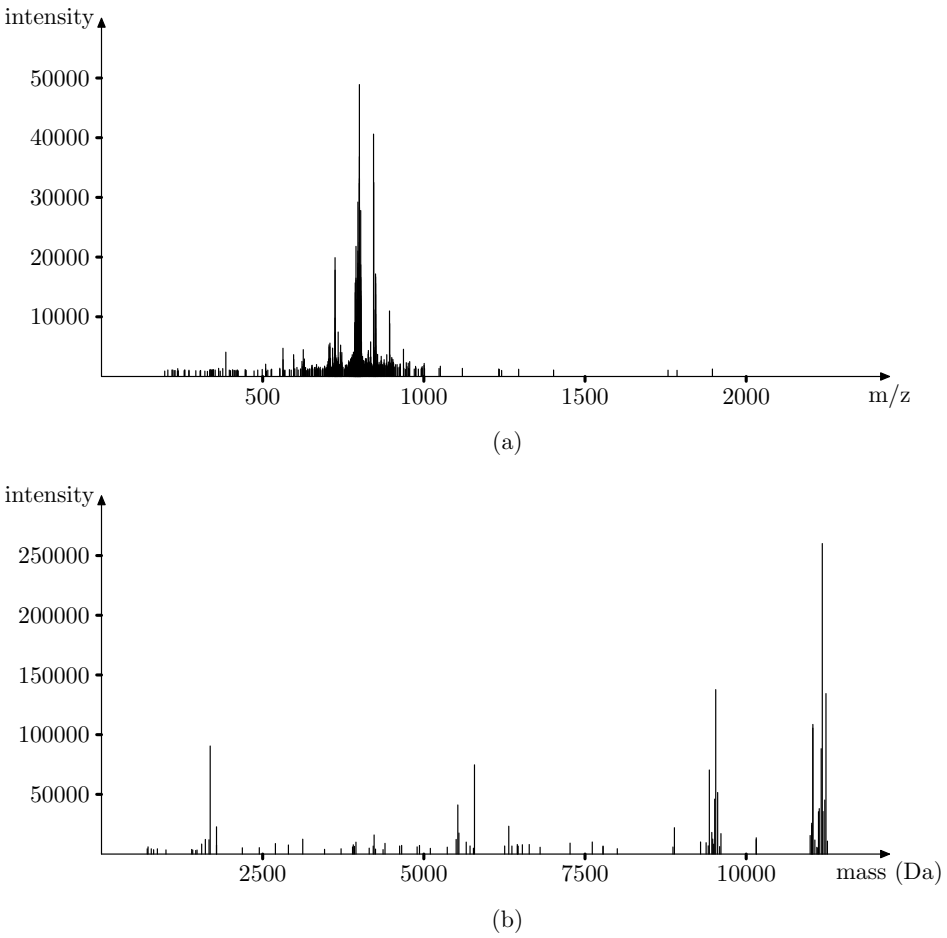

Figure 1.4: An example of top-down spectral deconvolution: (a) a centroided top-down mass spectrum; (b) a deconvoluted top-down mass spectrum of the spectrum in (a). Fragment ion peaks of various charge states and isotopomers are converted into neutral monoisotopic fragment masses.

Table 1.1: Software tools for proteoform identification using top-down mass spectrometry

Software	Website	Reference
ProSightPC	http://proteinaceous.net/product/prosightpc-4-0/	[27]
MS-TopDown	http://proteomics.ucsd.edu/software-tools/	[28]
PIITA	-	[29]
Mascot Top Down	http://www.matrixscience.com/	[30]
BUPID Top-Down	http://www.bumc.bu.edu/cardiovascularproteomics/cpctools/bupid-top-down/	[31]
MS-Align+	http://bix.ucsd.edu/projects/msalign/	[32]
Byonic	http://www.proteinmetrics.com/products/byonic/	[33]
MS-Align-E	http://proteomics.informatics.iupui.edu/software/msaligne/	[35]
ProteinGoggle	http://proteingoggle.tongji.edu.cn/	[34, 40]
ProSight Lite	http://prosightlite.northwestern.edu/	[41]
Proteoform Suite	https://github.com/smith-chem-wisc/ProteoformSuite	[39]
TopPIC	http://proteomics.informatics.iupui.edu/software/toppic/	[36]
pTop	http://pfind.ict.ac.cn/software/pTop/index.html	[37]
TopMG	http://proteomics.informatics.iupui.edu/software/topmg/	[1]
MASH Suite Pro	http://crb.wisc.edu/yinglab/software.html	[38, 42]
MSPathFinder	https://omics.pnl.gov/software/mspathfinder	[18]

Spectral alignment is used by many software tools to identify proteoforms with unexpected alterations [1, 28, 32, 35–37]. The spectral alignment algorithm finds an optimal alignment between the spectrum S and the protein sequence P by inserting 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+ [32] and TopPIC [36] are two 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 [35], MSPathFinder [18], and pTop [37], have been proposed to identify proteoforms with many variable PTMs.

One primary goal of translational research is to identify the molecular signatures or biomarkers of specific diseases or disease phenotypes from patient samples. After being found, these biomarkers often provide novel methods to detect and treat particular diseases. Recent findings suggest that mRNA abundance is only weakly correlated to the real protein expression levels [43]. Here we argue that intact proteoforms represent an efficient class of biomarkers. since they can recognize the real biological differences in samples. With increasing precision, top-down MS has become the method of choice to measure proteoforms in their intact states. Many efforts have been made to identify and quantify the disease-related proteoforms [44–49], including type II diabetes and myocardial dysfunction.

1.4 Contributions

The primary goal of this dissertation is to develop a complete pipeline for complex proteoform identification. In this dissertation, we propose several novel algorithms for computational problems in proteoform identification using top-down MS/MS. We apply the proposed algorithms over simulated and real top-down MS/MS data sets to evaluate their performance. We summarize the work in this dissertation as follows and present more details in the following chapters (Figure 1.5).

1.4.1 Approximate spectrum-based filtering algorithms

Protein sequence filtering is an indispensable step in proteome-level analyses because it is extremely slow to align thousands of mass spectra against thousands of protein sequences. Two kinds of methods are widely used: tag-based methods and unmodified protein fragment (UPF)-based methods. Tag-based methods depend on consecutive fragment ions in the query spectrum, and the efficiency is limited in top-down MS data due to the missing peaks. UPF-based methods achieved satisfactory performance in identifying unexpected alterations. However, they may fail in the sequence filtration when the target proteoform contains more than two variable PTMs and/or unexpected alterations.

To address the above problems, we propose two approximate spectrum-based filtering (ASF) algorithms that quickly find a small number of candidate proteins from a large proteome database for a query spectrum whose target proteoform has multiple variable PTMs. In the proposed ASF algorithms, the query spectrum is transformed into an approximate spectrum by removing variable PTMs in the match between the target reference sequence and the spectrum. Experiments on simulated and real data set demonstrated ASF algorithms outperformed existing ones on proteoforms with multiple variable PTMs.

1.4.2 Mass graph-based alignment algorithms

Although spectral alignment [28] achieved great success in identifying proteoforms with variable PTMs and unknown mass shifts, existing alignment algorithms have their limitations: MS-Align+ [32] and TopPIC [36] can identify proteoforms with at most two unknown mass shifts; MS-Align-E [35] and pTop [37] can identify proteoforms with variable PTMs, but not those with terminal truncations; MSPathFinder [18] can identify variable PTMs, but the identification of terminal truncations depends on high-quality sequence tags.

In this dissertation, we design a new data structure, called *mass graphs*, to represent all possible proteoforms generated from one reference sequence with multiple variable PTMs and/or terminal truncations. We also propose mass graph-based alignment algorithms to identify proteoforms with multiple variable PTMs and/or terminal truncations. High accuracy has been reported in experiments on simulated and real top-down MS data sets.

1.4.3 Statistical significance estimation

A fundamental problem in proteoform identification is to distinguish between correct and incorrect identifications. Many methods have been developed, but they have limitations in estimating the statistical significance of identified complex proteoforms: distribution fitting methods are computationally efficient, but they may fail in estimating extremely small p-values; the generating function method cannot handle identified proteoforms with more than two mass shifts; the Markov chain Monte Carlo (MCMC) method in MS-DPR was designed for bottom-up MS and cannot be directly used for top-down MS since it does not allow any PTMs.

In this dissertation, we propose a new MCMC method for estimating the statistical significance of identified complex proteoforms with multiple variable PTMs. We design a new Markov chain model to represent proteins for top-down spectral interpretation and use a greedy algorithm for quick estimation of the similarity score between the query spectrum and a protein with multiple variable PTMs. We use four top-down MS data sets to evaluate our new method and show its high discriminative capacity.

1.5 Organization of the dissertation

The rest of this dissertation work is organized as follows. In Chapter 2, we describe the data sets used in this study. In Chapter 3, we present two approximate spectrum-based protein sequence filtering algorithms in which both variable PTMs and unexpected alterations are considered. In Chapter 4, we describe mass graphs, which efficiently represent proteoforms with multiple variable PTMs and/or terminal truncations. Mass graph-based alignment algorithms are also proposed to identify complex proteoforms with variable PTMs and/or terminal truncations. In Chapter 5, an MCMC method is proposed to estimate the statistical significance of identified PrSMs. In Chapter 6, we summarize the work of this dissertation and discuss some future directions.

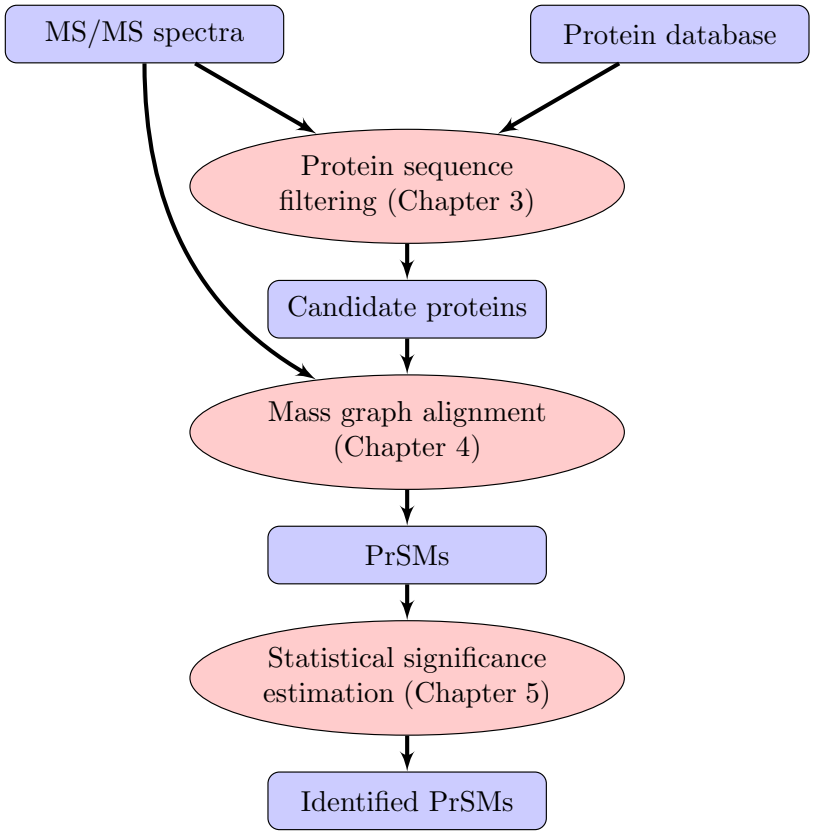

Figure 1.5: Overview of the dissertation work.

CHAPTER 2

DATA SETS

Five top-down MS data sets were used in this dissertation: the first was generated from *Escherichia coli* (EC) K-12 MG1655, the second and the third from purified human histone proteins, the fourth from breast tumor xenograft samples, and the fifth from human MCF-7 cells.

The EC data set was obtained using a LC system coupled with an LTQ Orbitrap Velos mass spectrometer (Thermo Scientific, Waltham, MA). The top 4 ions in each MS spectrum were selected for MS/MS analysis and the alternating fragmentation mode was used. With a resolution of 60 000, a total of 2 027 collision-induced dissociation (CID) and 2 027 electron-transfer dissociation (ETD) top-down MS/MS spectra were collected [36].

The first histone data set was generated from purified histone H4 protein [35]. 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. The histone H4 protein was further analyzed by an LTQ Orbitrap Velos mass spectrometer (Thermo Scientific, Waltham, MA). With a resolution of 60 000, a total of 1 626 CID and 1 626 ETD spectra were acquired.

The second histone data set was generated from the human histone H2A, H2B, H3, and H4 proteins. 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 by an LTQ Orbitrap Velos mass spectrometer (Thermo Scientific, Waltham, MA) with a 60 000 resolution. In total, 11 378 CID and 11 378 ETD top-down MS/MS spectra were collected, including 3 462 CID and 3 462 ETD spectra from histone H3 sample. More details of the MS experiment can be found in [50].

The breast tumor xenograft data set [51] was generated using an Orbitrap Elite mass spectrometer (Thermo Scientific, Waltham, MA). Cryopulverization of the tumor xenografts was performed using the standard CPTAC protocols [52]. A basal-like (WHIM2) breast cancer sample and a luminal B (WHIM16) breast cancer sample [53, 54] were used for

the experiments. Protein separation was achieved using a commercial GELFREE 8100 fractionation system (Expedeon, Cambridge, UK). With a resolution of 60 000, a total of 51 474 and 50 372 higher-energy collisional dissociation (HCD) top-down MS/MS spectra were collected from the WHIM2 and WHIM16 samples respectively.

For the MCF-7 data set, proteins extracted from MCF-7 cells were reduced with dithiothreitol and alkylated with iodoacetamide, and then separated by capillary zone electrophoresis (CZE). A one-meter linear polyacrylamide coated capillary (50 μm/360 μm i.d./o.d.) was used for CZE, and a commercialized electro-kinetically pumped sheath flow CE-MS interface (CMP Scientific, Brooklyn, NY) was used to couple CZE to MS [55, 56]. The background electrolyte (BGE) of CZE was 10% (v/v) acetic acid. The sample was dissolved in 50 mM ammonium bicarbonate (pH 8.0) for the dynamic pH junction based CZE-MS/MS [57], and injected into the capillary via applying 5-psi pressure for 95 seconds. The sample injection volume was 500 nL. 28 kV was applied across the capillary for separation and 2 kV was applied for electrospray. At the end of the separation, 20 psi was applied at the injection end for 10 min to flush the capillary with the BGE. A Q-Exactive HF mass spectrometer (Thermo Fisher Scientific, Waltham, MA) was coupled with the CZE system. The top 3 precursor ions in each MS spectrum were selected for MS/MS analysis. The mass resolution for MS and MS/MS was 120 000 and 60 000, respectively. The AGC target for MS and MS/MS was the same, 1E6. The number of microscan was 4 and 3 for MS and MS/MS, respectively. A total of 1 523 HCD MS/MS spectra were acquired.

All the raw data files were controided and converted to mzML files by msconvert in ProteoWizard (version 3.0.11537) [58]. The mzML files were further deconvoluted by TopFD (version 1.1.2), an improved version of MS-Deconv [23]. TopFD converted all MS/MS spectra into lists of neutral fragment masses. In TopFD, 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. TopFD often significantly simplifies top-down MS/MS spectra and converts a complex spectrum with thousands of peaks into a deconvoluted one with dozens or hundreds of fragment masses.

CHAPTER 3

PROTEIN SEQUENCE FILTERING ALGORITHMS FOR PROTEOFORM IDENTIFICATION

3.1 Introduction

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 PrSM [28]. It is extremely slow to align mass spectra against tens of thousands of database protein sequences [32]. 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 chapter, 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 [59]. In top-down MS, the method is employed in the Delta-M mode in ProSightPC [27]. 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 [60]. 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 [61–67]. In top-down MS, tag-based methods have been used in USTag [68], pTop [37], MSPathFinder [18], and the sequence tag mode in ProSightPC [27]. 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 [32,36]. 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 [69–71] have been proposed to partially solve the problem. In top-down MS, UPF-based methods have been used in MS-Align+ [32] and TopPIC [36] 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 are 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 chapter, we propose two *Approximate Spectrum-based 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 (Figure 3.1), significantly increasing filtering efficiency.

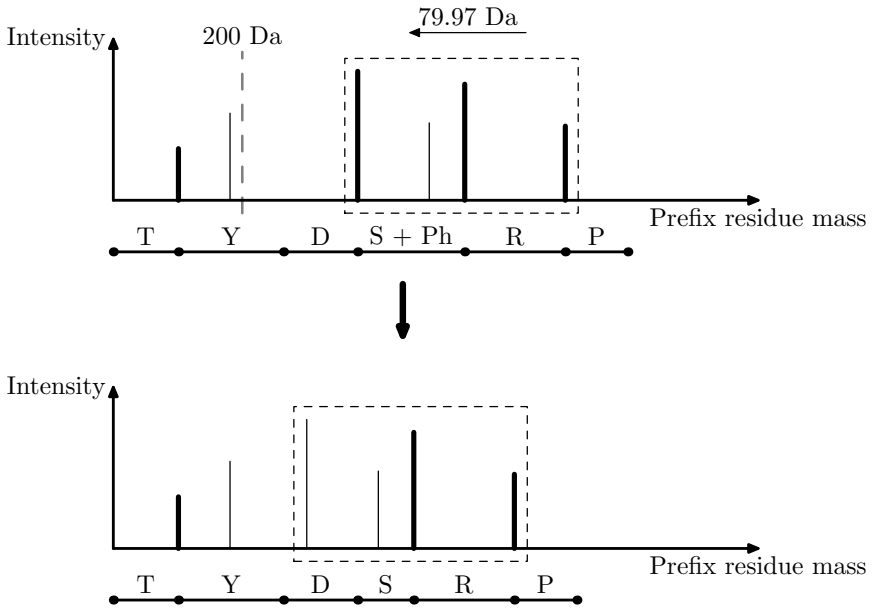

Figure 3.1: A prefix residue mass spectrum (top) of the proteoform TYDS[Ph]RP with a phosphorylation site on the serine residue is transformed into an approximate prefix residue mass spectrum (bottom) of the unmodified protein TYDSPR. In the top spectrum, each peak represents a possible prefix residue 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.

We evaluated the ASF algorithms and 4 existing ones for protein sequence filtration in top-down MS database search. Experiments on simulated data showed that the ASF algorithms outperformed the existing ones for complex proteoform identification. By combining the ASF and mass graph alignment algorithms [1], we identified many phosphorylated proteoforms missed by ProSightPC from a top-down MS data set of breast cancer xenograft samples.

3.2 Methods

In the ASF 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 tag-based and UPF-based filtering algorithms and then describe the ASF algorithms.

3.2.1 Tag-based filtering algorithms

A sequence tag is a short amino acid sequence extracted from an MS/MS spectrum. Most tag extraction methods are based on spectrum graphs [61]. A spectrum graph is constructed from a deconvoluted MS/MS spectrum using three steps (Figure 3.2): (a) A node is added to the spectrum graph for each fragment mass in the spectrum. (b) Two nodes are connected by an edge if the difference between their corresponding masses is similar to (within an error tolerance) the mass of an amino acid residue. In some tag generation methods, two nodes are connected if their corresponding mass difference is similar to the mass of one or two amino acid residues. The label of the edge is the amino acid. (c) A node is removed from the graph if there are no edges connecting to it. Each path in the spectrum graph corresponds to a sequence tag. A top-down spectrum graph typically consists of several connected components because of many missing peaks.

We describe two sequence tag-based filtering methods, which are used in MS-Align+Tag and MSPathFinder [18], respectively. The first method uses the long tag strategy to obtain sequence tags from a spectrum graph with three steps: (a) A longest sequence tag is selected from each component of the spectrum graph. If a component contains several longest sequence tags with the same length, one of them is arbitrarily selected. (b) The reported sequence tags are filtered to remove those with less than k amino acids ($k = 4$ in the experiments). (c) For each remaining sequence tag, all of its substrings with length k are reported. For example, in Figure 3.2, the longest sequence tags NVYTSAG and AC are extracted from the spectrum graph, then the tag AC is filtered out because its length is less than $k = 4$, and finally four length-4 short tags are extracted: NVYT, VYTS, YTSA, and TSAG.

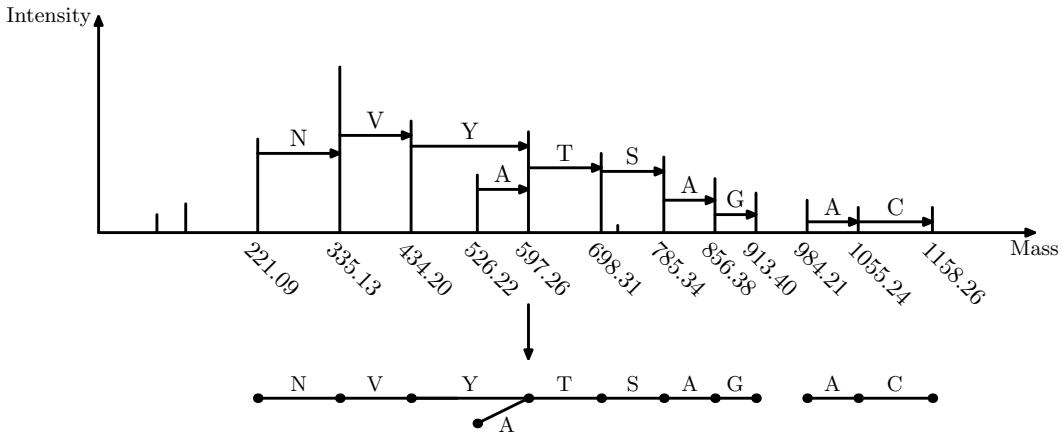

Figure 3.2: A spectrum graph (bottom) is constructed from a deconvoluted MS/MS spectrum (top). The two left most nodes correspond to masses 221.09 Da and 335.13 Da in the spectrum. These two nodes are connected by an edge because the difference between 221.09 and 335.13 is similar to the mass of an asparagine residue (114.04 Da). The spectrum graph contains two connected components.

In the second method, we extract from the spectrum graph all sequence tags with a length l between the minimum length l_{min} and the maximum length l_{max}, that is, $l_{min} \leq l \leq l_{max}$. In the experiment, $l_{min} = 5$ and $l_{max} = 8$. First, all tags with length l_{max} are extracted from the spectrum graph and added to a sequence tag set T. For example, when $l_{max} = 6$, two tags NVYTSA and VYTSAG are extracted from the graph in Figure 3.2. Next, all tags with length $l_{max} - 1$ are extracted. A length $l_{max} - 1$ tag is added to T if it is not a substring of any tag in T. For example, the length-5 sequence tag NVYTS in Figure 3.2 is not added to T because it is a substring of the length-6 sequence tag NVYTSA, and the sequence tag ATSAG is added to T because it is not a substring of any tag in T. Two tags in T may share a substring, but their whole sequences are different. Similarly, we further extract sequence tags with lengths $l_{max} - 2, \ldots, l_{min}$ and add them to T if they are not substrings of tags in T. The two methods are called TAG-LONG (with the long tag strategy) and TAG-VAR (with tags of various lengths), respectively.

Because some sequence tags are extracted from suffix fragment ion series, a reversed tag is generated from each extracted tag. The extracted sequence tags and their reversed tags are searched against a protein database to find a small number of top candidate proteins. Because the lengths of proteins vary significantly from dozens to tens of thousands, we compute similarity scores between sequence tags and protein fragments with similar lengths

rather than whole proteins. Protein fragments are generated using a parameter L ($L = 150$ in the experiments). If the length of a protein is no larger than L, the whole protein sequence is a fragment. Otherwise, each length L substring of the protein is a fragment, and the total number of fragments of the protein is $n - L + 1$.

Let T be a set of sequence tags and reversed tags extracted from a spectrum graph. We define a similarity score between a candidate fragment and T. If a sequence tag is a substring of a fragment, we say the sequence tag has a hit in the fragment. The *tag score* between the fragment and T is the number of tags in T that have a hit in the fragment. The *tag score* between a protein and T is the maximum tag score among its fragments. All proteins in the protein database are ranked based on their tag scores and the top t ($t = 20$ in experiments) proteins are reported as filtering results.

3.2.2 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 $P = a_1 a_2 \dots a_n$ is the sum of the residue masses of its amino acids, that is, $\sum_{k=1}^{n} \text{mass}(a_k)$. The residue mass of the length-i prefix $a_1 a_2 \dots a_i$ is a prefix residue mass of P, denoted by p_i. The residue mass of the length-i suffix $a_{n-i+1} \dots a_n$ 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 $P_{pre} = \{p_0, p_1, \dots, p_n\}$ and the set of all suffix residue masses of P as $P_{suf} = \{s_0, s_1, \dots, s_n\}$.

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 CID spectrum, both the prefix residue mass set and the suffix residue mass set contain the following two masses: 0 and $M - \text{mass}(H_2O)$, 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 $x - \text{mass}(H_2O)$ and $M - x - \text{mass}(H_2O)$ 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 $M - \text{mass}(H_2O) = 284.16$ are added into S_{pre} and S_{suf}. $S_{pre} = \{0, 71.04, 128.06, 174.11, 231.13, 284.16\}$ after the masses x and $M - x$ for fragment masses x are added; $S_{suf} = \{0, 53.03, 110.05, 156.10, 213.12, 284.16\}$ after the masses $x - \text{mass}(H_2O)$ and $M - x - \text{mass}(H_2O)$ 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, we choose c, z-dot, and z-prime ions as the most commonly observed ones in ETD spectra, and 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 [36]. 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 (Figure 3.3(a)), denoted by $D(A, B) = \max_d C(\text{shift}(A, d), B)$. 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 Figure 3.3(a) contains an unmodified fragment IDES. When a CID spectrum of T[Ph]IDEST[Ph]R contains peaks of the b-ions b_1, $b_2, \ldots,$ 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 (Figure 3.3(b)), denoted by $R(A, B) = \max_{d \in A} C(\text{shift}(A, -d), B)$.

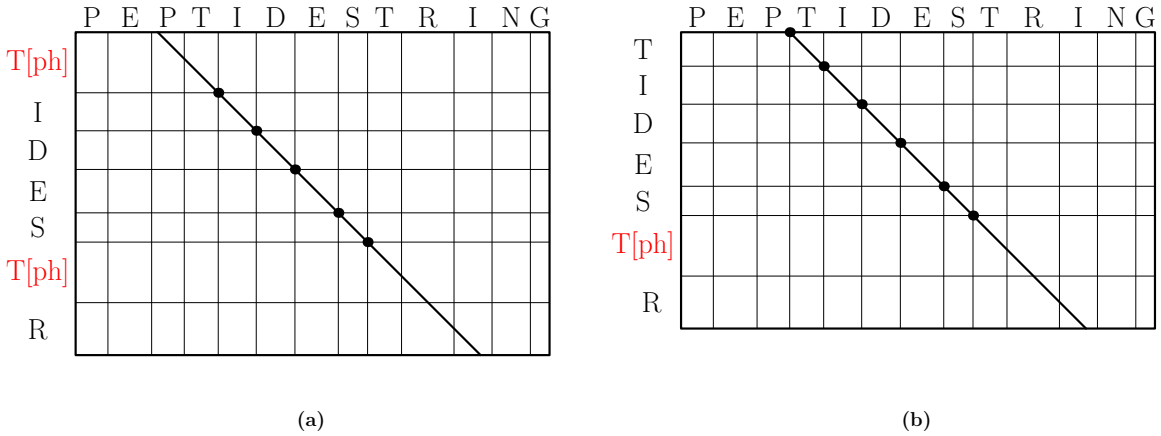

Figure 3.3: Diagonal scores and restricted diagonal scores. (a) The diagonal score between the prefix residue masses of PEPTIDESTRING and T[Ph]IDEST[Ph]R is 5, corresponding to the 5 dots in the diagonal. The score is obtained by shifting the prefix residue masses of PEPTIDESTRING by -243.18 Da, which equals $-\text{mass(PEPT)} + \text{mass(T[Ph])}$. (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 $= -\text{mass(PEP)}$.

For example, when A is the set of prefix residue masses $\{0, 97.05, 226.09\}$ of the peptide PE, $R(A, B) = \max\{C(\text{shift}(A, 0), B), C(\text{shift}(A, -97.05), B), C(\text{shift}(A, -226.09), B)\}$. 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 Figure 3.3(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 computation of a 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 a high 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 in filtration. 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 [69]. The two methods are called UPF-DIAGONAL (the diagonal score) and UPF-RESTRICT (the restricted diagonal score), respectively.

3.2.3 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 3.4). 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 $P = a_1 \dots a_i \dots a_n$ 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 $P_{pre} = \{p_0, p_1, \dots, p_i, p_{i+1} \dots, p_n\}$ contains all prefix residue masses of P and the theoretical spectrum F_{pre} contains all prefix residue masses of F, that is, $F_{pre} = \{p_0, p_1, \dots, p_i+\delta, p_{i+1}+\delta, \dots, p_n+\delta\}$. We can convert F_{pre} into P_{pre} by deducting δ from the prefix residue masses $p_i+\delta, p_{i+1}+\delta, \dots, p_n+\delta$.

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, $S_{pre} = F_{pre} = \{p_0, p_1, \dots, p_i+\delta, p_{i+1}+\delta, \dots, p_n+\delta\}$. In the ASF method, we try to convert S_{pre} into an approximate spectrum of P_{pre} with limited information (Figure 3.1): 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 $M-\text{mass}(H_2O)$. 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 $\text{conv}(S_{pre},q)$ by deducting δ from all masses in S_{pre} that are no less than q. In Figure 3.1, 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 Figure 3.1, 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 $\text{conv}(S_{pre},q)$. 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 $\text{conv}(S_{pre},q)$ 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 $\frac{p_n+\delta}{2k}$, and the conversion error ratio of $\text{conv}(S_{pre},q^*)$ is no larger than $\frac{1}{2k}$. When k is large, $\text{conv}(S_{pre},q^*)$ 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 $(kf)^h$, where f is the number of variable PTM types in database search. 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 (Figure 3.4). Detailed steps for Step 4 in the algorithm is given in Figure 3.5. 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 3.3.2), and h was set as 1 or 2. The number f of variable PTM types is a parameter specified by the user.

3.3 Results

3.3.1 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 [72] (version September 12, 2016, 4 306 entries) and concatenated with a shuffled decoy database of the same size. The 4 054 top-down MS/MS spectra were deconvoluted by TopFD and then searched against the target-decoy concatenated EC proteome database using TopPIC [36].

The ASF-RESTRICT algorithm

Input: A deconvoluted top-down MS/MS spectrum S, a set Ω of f variable PTMs, a number k of intervals, parameters h and t, and a protein database D.

Output: Top t candidate protein sequences in D for the query spectrum S.

1. Set the protein set Φ as an empty set, and compute k intervals as well as their k centers in S.
2. **For** each set of h masses selected from the k centers with replacement **do**
3. **For** each set of h PTMs selected from Ω with replacement **do**
4. Generate an approximate spectrum S' using the h selected masses and the h selected PTMs.
5. Use the UPF-RESTRICT algorithm to search S' against D to find top t candidate proteins as well as their similarity scores, and add them to Φ.
6. Report t top scoring protein sequences from Φ.

Figure 3.4: The ASF-RESTRICT algorithm for protein sequence filtration using top-down MS/MS spectra.

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).

For each identified PrSM between a spectrum S and a protein sequence P with a score x, we used the generating function method [73, 74] 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 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 approximate spectrum generation algorithm

Input: A deconvoluted top-down MS/MS spectrum S with a precursor mass M and peaks $(a_1, b_1), (a_2, b_2), \ldots, (a_n, b_n)$, where a_i is the ith mass and b_i is the intensity of a_i; h guessed prefix residue masses $c_1 \leq c_2 \leq \ldots c_h$; and h guessed PTMs and their corresponding mass shifts $\delta_1, \delta_2, \ldots, \delta_h$.

Output: An approximate spectrum S'.

1. Set $q_0 = 0$, $q_{h+1} = M$, and $q_k = c_k$ for $1 \leq k \leq h$.
2. **For** $i = 1$ to n **do**
3. Find two values q_j and q_{j+1} such that $q_j \leq a_i < q_{j+1}$.
4. $a'_i = a_i - \sum_{k=1}^{j} \delta_k$.
5. **If** $a'_i > 0$ **then** add (a'_i, b_i) as a peak to S'.
6. Set the precursor mass of S' as $M - \sum_{k=1}^{h} \delta_k$ and output S'.

Figure 3.5: An algorithm for generating an approximate spectrum from a query top-down deconvoluted MS/MS spectrum S and a list of guessed prefix residue masses and variable PTMs.

the resulting sequence and S contains a PTM (mutation), an N-terminal truncation, and a C-terminal truncation. Using this method, a total of 13 110 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 65 550 PrSMs (13 110 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.

3.3.2 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 of the ASF algorithms with various settings for $k = 2, 3, 4, 5, 6$ and $h = 1, 2$ are shown in Figure 3.6. 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.

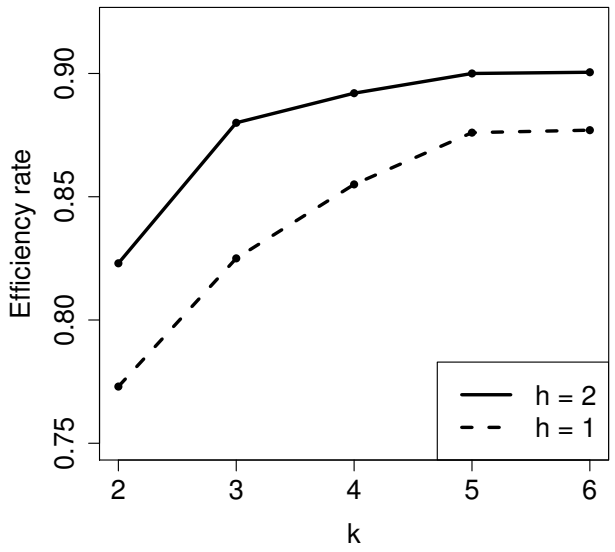

Figure 3.6: The efficiency rates of the ASF algorithms with various settings $k = 2, 3, 4, 5, 6$ and $h = 1, 2$ on the simulated PrSMs with 5 PTMs.

3.3.3 Evaluation on filtration efficiency

We tested the TAG-LONG, TAG-VAR, UPF-RESTRICT, UPF-DIAGONAL, ASF-RESTRICT and ASF-DIAGONAL algorithms on the simulated PrSMs with 5 PTMs. 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%. The ASF-DIAGONAL algorithm missed 528, 253, and 794 PrSMs efficiently filtered by UPF-RESTRICT, UPF-DIAGONAL, and ASF-RESTRICT, respectively.

The efficiency rates of the filtering algorithms are related to the conditional spectral probabilities of test PrSMs (Figure 3.7). Most PrSMs with a conditional spectral probability $\geq 10^{-30}$ 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.

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 fragment. 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 fragment. The experimental results show that combining the two methods can improve filtration efficiency.

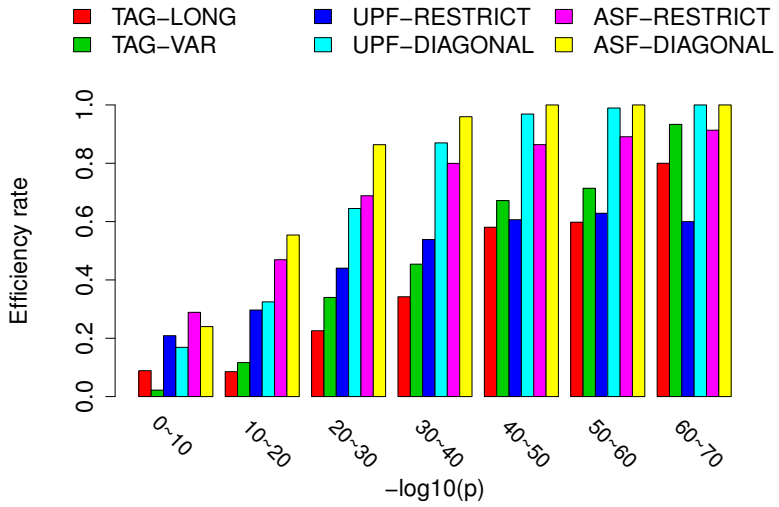

Figure 3.7: Comparison of the filtration efficiency rates of the TAG-LONG, TAG-VAR, UPF-RESTRICT, UPF-DIAGONAL, ASF-RESTRICT and ASF-DIAGONAL algorithms on the simulated test PrSMs with 5 PTMs. The PrSMs are divided into 7 groups based on their conditional spectral probabilities p, and the efficiency rates for each group are compared.

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 [72] (version July 9, 2016, 20 191 entries). The concatenated database contained 24 497 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%.

3.3.4 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. TopMG [1] 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 3.1) were used as variable PTMs in proteoform identification. TopMG identified 3 205 and 1 087 PrSMs with at least 10 matched fragment ions from the histone H3 and H4 data sets, respectively.

Table 3.1: Five variable PTMs used in the identification of proteoforms of histone proteins

PTM	Monoisotopic mass shift (Da)	Amino acids
Acetylation	42.01056	R, K
Methylation	14.01565	R, K
Dimethylation	28.03130	R, K
Trimethylation	42.04695	R
Phosphorylation	79.96633	S, T, Y

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 July 9, 2016, 20 191 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.

The filtration efficiency rates of the 6 filtering methods for the histone H3 and H4 PrSMs are summarized in Table 3.2. 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 ASF-RESTRICT were the highest among the 6 methods. Most of the histone H3 and H4 proteoforms have no more than 4 PTMs, 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, ASF-RESTRICT 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 UPF-RESTRICT and 1 094 histone H3 PrSMs missed by ASF-DIAGONAL. Similarly, ASF-RESTRICT outperformed ASF-DIAGONAL and UPF-RESTRICT on the histone H4 PrSMs. 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. 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 (Figure 3.7).

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.

3.3.5 Phosphorylated proteoforms identified from the xenograft data set

The ASF algorithms were combined with TopMG [1] 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 [27] 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 TopFD. 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 [36] 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 [51], 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 November 13, 2016, 16 840 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. 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. 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.

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. 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 TopFD with an error tolerance 2.2 Da to find candidate PrSMs. Of the 20 proteoforms missed by TopPIC, TopFD 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. Of the remaining 23 phosphorylated proteoforms (14 precursor masses) reported by ProSightPC, 4 did not have matched deconvoluted spectra reported by TopFD, 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}.

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, 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.

In the second step of the analysis of TopPIC and TopMG, the human proteome database (version July 9, 2016, 20 191 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. Similar to the first step, most of the identified phosphorylated proteoforms contain ≤ 3 phosphorylation sites.

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. 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. The proteoforms identified by TopMG include 47 proteoforms with an E-value smaller than 10^{-10}. Similar to the comparison on mouse phosphorylated proteoforms, TopMG identified many phosphorylated human proteoforms missed by the absolute mass and biomarker modes of ProSightPC.

3.4 Discussion

In this chapter, 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 variable 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 is 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.

Table 3.2: Comparison of the 6 filtering algorithms in the filtration efficiency rate using the 3 205 histone H3 PrSMs and the 1 087 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

CHAPTER 4

MASS GRAPH ALIGNMENT

4.1 Introduction

Extended proteoform databases and spectral alignment are the two main database search strategies for proteoform identification. ProSightPC [27] and MascotTD [30] use the first approach, in which spectra are searched against a sequence database of commonly observed proteoforms. However, the number of candidate proteoforms increases exponentially due to the combinatorial explosion of PTMs and truncations. As a result, most uncommon proteoforms have to be excluded from the sequence database to keep its size manageable, limiting the ability to identify uncommon or novel proteoforms.

Spectral alignment [28] is capable of identifying variable PTMs and unknown mass shifts by finding a best scoring alignment between the spectrum and the reference sequence. However, existing alignment algorithms have their limitations. MS-Align+ [32] and TopPIC [36] can identify proteoforms with at most two unknown mass shifts because it treats all PSAs as unknown mass shifts except for fixed PTMs and protein N-terminal PTMs. MS-Align-E [35] and pTop [37] are capable of identifying proteoforms with variable PTMs, but not those with terminal truncations. MSPathFinder [18] is also capable of identifying variable PTMs, but the identification of truncations depends on high quality sequence tags.

In this chapter, we use *mass graphs* (Figure 4.1) to efficiently represent proteoforms of a protein with variable PTMs and/or terminal truncations. We transform the proteoform identification problem to the mass graph alignment problem and propose dynamic programming algorithms for a restricted version of the problem.

Many graph-based approaches have been proposed in bioinformatics studies. Splicing graphs were proposed by Heber et al. [75] for solving the EST assembly problem and have been widely used in the identification of alternative splicing events [76]. In proteogenomics studies, splicing graphs [77] and variant graphs [78] were employed for representing transcript variants. In the variant graph approach, both genetic variations and alternative splicing junctions of a gene are represented in a variant graph, in which each node rep-

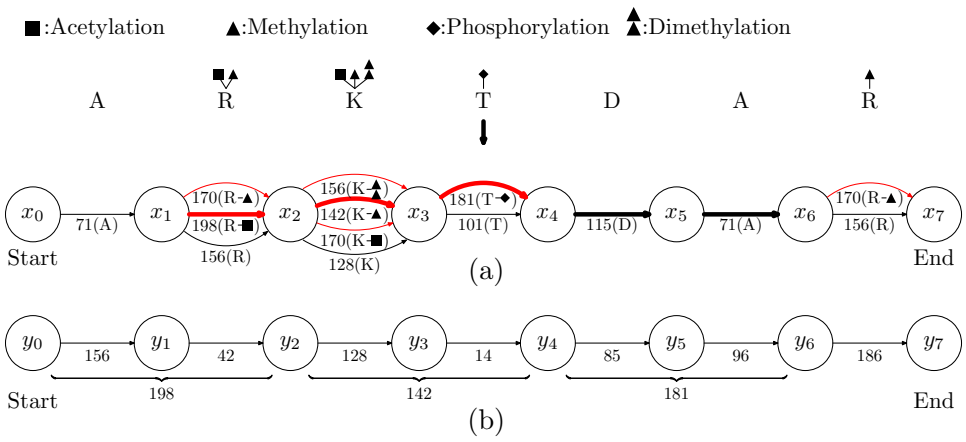

Figure 4.1: Construction of mass graphs. (a) An illustration of the construction of a proteoform mass graph from a protein ARKTDAR and four variable PTMs: acetylation on K and the first R; methylation on R and K, phosphorylation on T, and dimethylation on K. Each node corresponds to a peptide bond, or the N- or C-terminus of the protein; each edge corresponds to an amino acid residue (red edges correspond to modified amino acid residues). The weight of each edge is the mass of its corresponding unmodified or modified residue (a scaling factor 1 is used to convert weights to integers). (b) An illustration of the construction of a spectral mass graph from a prefix residue mass spectrum $0, 156, 198, 326, 340, 425, 521, 707$. The spectrum is generated from a proteoform of RKTDA with an acetylation on the R, a methylation on the K, and a phosphorylation on the T. To simplify the mass graph, masses corresponding to proteoform suffixes (C-terminal fragment masses) are not shown. The full path from the start node y_0 to the end node y_7 is aligned with the bold path from node x_1 to node x_6. The path from y_0 to y_6 and the red bold path from x_1 to x_4 are consistent.

resents a sequence of nucleotide bases and each path corresponds to a transcript variant of the gene. The transcript variants represented in a variant graph are translated into peptide or protein sequences for the identification of MS/MS spectra. Splicing graphs and variant graphs efficiently represent an exponential number of transcript variants and their corresponding proteoforms. Another example of graph-based methods is spectrum graphs that were proposed for *de novo* peptide sequencing and sequence tag generation in MS data analysis [61, 79]. In a spectrum graph, each node represents a prefix residue mass in an MS/MS spectrum, and each path represents a peptide that may explain the spectrum. He et al. [80] extended the spectrum graph approach to incorporate limited number of PTMs, and Bhatia et al. [81] proposed to use a constraint graph to represent sequence constraints and combine a spectrum graph and a constraint graph in *de novo* sequencing.

The idea of mass graphs is inspired by splicing graphs, variant graphs, spectrum graphs, and constraint graphs. Similar to variant graphs, a mass graph efficiently represents an exponential number of possible proteoforms of a gene. In addition, mass graphs are capable of representing site specific variable PTMs. Compared with variant graphs and spectrum graphs, the mass graph representation has its unique properties. While variant graphs store sequences of nucleotide bases (which can be translated into amino acids sequences) in nodes, mass graphs store amino acid residue masses in edges. Replacing nucleotides (or amino acids) with masses simplifies the representation of proteoforms with variable PTMs. (See Section 4.4.) While nodes in a spectrum graph represent prefix residue masses of an MS/MS spectrum, nodes in a mass graph represent prefix residue masses of many possible proteoforms.

The mass graph alignment problem is different from the spectral alignment problem [28, 82] and the spliced alignment problem [83]. While spectral alignment methods search for the best alignment between *two* lists of prefix residue masses, the mass graph alignment problem finds the best alignment between a prefix residue mass list and all possible paths in a mass graph, each of which corresponds a prefix residue mass list and a proteoform. In the spliced alignment problem, a variation of a nucleotide base does not significantly affect the whole sequence alignment. However, a mass shift in an amino acid and its corresponding edge in a mass graph dramatically affect the similarity score between a prefix residue mass list and a path containing the edge because the mass shift "propagates" to the residue masses of all prefixes containing the amino acid. (See Section 4.4.)

We propose TopMG (TOP-down mass spectrometry-based proteoform identification using Mass Graphs), a software tool for identifying modified proteoforms using top-down tandem mass spectra, which is based on algorithms for the mass graph alignment problem. TopMG was tested on three top-down MS/MS data sets. Experimental results showed that TopMG was efficient in identifying proteoforms with variable PTMs and outperformed MS-Align-E [35] and ProSightPC [27] in identifying complex proteoforms, especially those with terminal truncations.

4.2 Methods

Mass graphs are used to represent candidate proteoforms and top-down MS/MS spectra. Mass graphs representing proteoforms are called *proteoform mass graphs*; those representing MS/MS spectra *spectral mass graphs*. With the representation, we formulate the proteoform identification problem as the mass graph alignment problem and design dynamic programming algorithms for a restricted version of the problem.

4.2.1 The mass graph alignment problem

Proteoform mass graphs A proteoform mass graph is constructed from an unmodified protein sequence and its variable PTMs with three steps (Figure 4.1(a)). (1) A node is added to the graph for each peptide bond of the protein. In addition, a start node and an end node are added for the N and C-termini of the protein, respectively. The *left node* of an amino acid is the one representing the peptide bond left of the amino acid. Specifically, the start node is the left node of the amino acid at the N-terminus. The *right node* of an amino acid is the one representing the peptide bond right of the amino acid. Specifically, the end node is the right node of the amino acid at the C-terminus. (2) For each amino acid in the protein, we add into the graph a directed black edge from its left node to its right node. The weight of the edge is the residue mass of the amino acid. (3) If an amino acid is a site of a variable PTM, we add into the graph a directed red edge from its left node to its right node. The weight of the edge is the residue mass of the amino acid with the PTM.

The locations of a PTM can be specified in a mass graph, thus reducing the number of candidate proteoforms. For example, the mass graph in Figure 4.1(a) specifies that acetylation occurs on only the first arginine residue, not the second, in the protein. As a result, mass graphs are capable of representing amino acid mutations because a mutation can be treated as a variable PTM that modifies only the amino acid at the mutation site. To represent an amino acid with a fixed PTM, the weight of the black edge corresponding to the amino acid is assigned as the mass of the residue with the fixed PTM.

Each path in a mass graph represents a proteoform of the protein. A path from the start node to the end node is called a *full path* of the graph, representing a proteoform

without terminal truncations. In the graph, the number of nodes is proportional to n, and the number of edges is proportional to ln, where n is the length of the protein sequence and l is the largest number of edges between two nodes.

Spectral mass graphs Mass graphs are also used to represent top-down MS/MS spectra. In the preprocessing of spectra, peaks are converted into neutral monoisotopic masses of fragment ions by deconvolution algorithms [19, 23, 25]. Peak intensities are ignored to simplify the description of the methods. These monoisotopic masses are further converted to a list of candidate prefix residue masses, called a prefix residue mass spectrum [35].

A prefix residue mass spectrum with masses $a_0, a_1, \ldots, a_n$ in the increasing order is converted into a spectral mass graph as follows (Figure 4.1(b)). A node is added into the graph for each mass in the spectrum. The nodes for $a_0 = 0$ and $a_n = PrecMass -$ mass(H_2O) are labeled as the start and the end nodes, respectively. For each pair of neighboring masses a_i and a_{i+1}, for $0 \leq i \leq n-1$, a directed edge is added from the node of a_i to that of a_{i+1}, and the weight of the edge is $a_{i+1} - a_i$. The spectral mass graph contains only one full path.

In the construction of mass graphs, the masses of all amino acids and PTMs are scaled and rounded to integers (a scaling constant 274.335215 was used in the experiments [35]). Precursor masses and candidate prefix residue masses in highly accurate top-down mass spectra are discretized using the same method. As a result, all edge weights are integers in mass graphs.

Formulation of the mass graph alignment problem With the mass graph representation, the proteoform identification problem is transformed to an alignment problem between a proteoform mass graph and a spectral mass graph. The objective of the alignment problem is to find a path in the spectral mass graph and a path in the proteoform mass graph such that the similarity score between the two paths is maximized.

Let A be a path with k edges $e_1, e_2, \ldots, e_k$. The weight of the prefix $e_1, e_2, \ldots, e_i$, $1 \leq i \leq k$, is called a prefix weight of A, denoted as w_i. Specifically, $w_0 = 0$ and w_k is the weight of the whole path. The path A is also represented as a list of prefix weights

$w_0, w_1, \ldots, w_k$. For example, the prefix weight list of the red bold path in Figure 4.1(a) is $0, 198, 340, 521$. Two paths are *consistent* if their weights are the same. For example, the red bold path in Figure 4.1(a) and the path from y_0 to y_6 in Figure 4.1(b) are consistent because they have the same weight 521.

We define the shared mass counting score of two consistent paths A and B as the number of shared prefix weights in their prefix weight lists, denoted as Score(A, B). For example, the shared mass counting score of the red bold path in Figure 4.1(a) and the path from y_0 to y_6 in Figure 4.1(b) is 4 because they share 4 prefix masses 0, 198, 340, and 521. If A and B are inconsistent, Score$(A, B) = -\infty$.

Given a proteoform mass graph G and a spectral mass graph H, the *mass graph alignment problem* is to find a path A in G and a path B in H such that Score(A, B) is maximized. There are several variants of the mass graph alignment problem. In the local alignment problem, the two paths in the mass graphs are not required to be full paths (from the start to the end node). It can identify a sequence tag of the target proteoform as well as its matched masses in the spectrum. For example, the alignment between the red bold path in Figure 4.1(a) and the path from y_0 to y_6 in Figure 4.1(b) is a local alignment. The proteoform identification problem is transformed into the semi-global mass graph alignment problem in which the path B in the spectral mass graph is required to be the full path. If the path A is a full path, a proteoform without terminal truncations is identified. Otherwise, a truncated proteoform is reported. For example, the bold path (not a full path) from x_1 to x_6 in Figure 4.1(a) is aligned with the full path in Figure 4.1(b), corresponding to a truncated proteoform R[Acetylation]K[Methylation]T[Phosphorylation]DA. In the global alignment problem, both A and B are required to be full paths, that is, terminal truncations are forbidden.

In proteoform identification, we can reduce the search space by limiting the number of PTM sites in a proteoform. This limitation gives rise to a variant of the mass graph alignment problem in which the number of red edges corresponding to modified amino acids is limited. Given a proteoform mass graph G, a spectral mass graph H, and a number t, the *restricted mass graph alignment (RMGA) problem* is to find a path A in G and a path B in H such that A contains no more than t red edges and Score(A, B) is maximized.

The algorithm for computing all the r-distance sets

Input: A proteoform mass graph G with nodes $x_0, x_i, \ldots, x_n$ in the topological order, and a number t.

Output: The distance sets $D(x_i, x_j, r)$ for $0 \leq i \leq j \leq n$ and $0 \leq r \leq t$.

```
1.  For i = 0 to n do
2.      Set D(x_i, x_i, 0) = {0} and set D(x_i, x_i, r) = ∅ for 1 ≤ r ≤ t.
3.  For i = 0 to n do
4.      For j = i + 1 to n do
5.          For r = 0 to t do
6.              Initialize D(x_i, x_j, r) = ∅.
7.              If r ≥ 1 then
8.                  For each red edge e_r ∈ R(x_{j-1}, x_j) do
9.                      For each d ∈ D(x_i, x_{j-1}, r − 1) do
10.                         Add d + w(e_r) into D(x_i, x_j, r).
11.             For each black edge e_b ∈ B(x_{j-1}, x_j) do
12.                 For each d ∈ D(x_i, x_{j-1}, r) do
13.                     Add d + w(e_b) into D(x_i, x_j, r).
```

Figure 4.2: The algorithm for computing all the r-distance sets of a proteoform mass graph.

4.2.2 Consistent preceding node pairs

We use consistent preceding node pairs described below to solve the RMGA problem. In a mass graph, if there is a path from a node u_1 to another node u_2, we say u_1 precedes u_2. There may exist different paths from u_1 to u_2, each of which defines a distance that equals the weight of the path. Let $D(u_1, u_2)$ denote the set of all distinct distances defined by the paths from u_1 to u_2. The size of $D(u_1, u_2)$ is smaller than the number of paths from u_1 to u_2 when there are many duplicated distances introduced by consistent paths. For example, in Figure 4.1(a), there are a total of 12 paths from x_1 to x_3, but $D(x_1, x_3)$ contains only 7 distances $\{284, 298, 312, 326, 340, 354, 368\}$. When u_1 is not a preceding node of u_2, $D(u_1, u_2)$ is an empty set.

Let u_1, u_2 be two nodes in G and let v_1, v_2 be two nodes in H. The node pair (u_1, v_1) is a consistent preceding node pair of the other node pair (u_2, v_2) if $D(u_1, u_2) \cap D(v_1, v_2) \neq \emptyset$, that is, there exist two consistent paths: one from u_1 to u_2, the other from v_1 to v_2. For example, the node pair (x_1, y_0) is a consistent preceding node pair of the node pair (x_3, y_4) in Figure 4.1, because $D(x_1, x_3) \cap D(y_0, y_4) = \{340\}$.

Given a proteoform mass graph G and a spectral mass graph H, the *consistent preceding node pair problem* is to find all consistent preceding node pairs for every node pair (u, v) where u is in G and v is in H. We study a variant of the problem in which the number of red edges in a path in G is restricted. Let $D(u_1, u_2, r)$ denote the set of distances defined by the paths from u_1 to u_2 that contain exactly r red edges, called an r-distance set. A node pair (u_1, v_1) is an r-consistent preceding node pair of the other node pair (u_2, v_2) if $D(u_1, u_2, r) \cap D(v_1, v_2) \neq \emptyset$.

Computing r-distance sets Let $x_0, x_1, \ldots, x_n$ be the nodes in the proteoform mass graph G in the topological order. We propose a dynamic programming algorithm (Figure 4.2) for computing $D(x_i, x_j, r)$ for $0 \leq i \leq j \leq n$ and $0 \leq r \leq t$. In the initialization (Steps 1 and 2), we set for each node x_i in G

$$D(x_i, x_i, r) = \begin{cases} \{0\} & \text{if } r = 0; \\ \emptyset & \text{otherwise.} \end{cases}$$

For $0 \leq i < j \leq n$ and $0 \leq r \leq t$, the set $D(x_i, x_j, r)$ is computed based on the distances between x_i and x_{j-1}. Let $R(u_1, u_2)$ ($B(u_1, u_2)$) be the set of all red (black) directed edges from a node u_1 to another node u_2. The weight of an edge e is denoted by $w(e)$. For each red edge $e_r \in R(x_{j-1}, x_j)$ and each distance $d \in D(x_i, x_{j-1}, r-1)$, we add $d + w(e_r)$ into $D(x_i, x_j, r)$ (Steps 7-10). For each black edge $e_b \in B(x_{j-1}, x_j)$ and each distance $d \in D(x_i, x_{j-1}, r)$, we add $d + w(e_b)$ into $D(x_i, x_j, r)$ (Steps 11-13). When the number of the types of variable PTMs in proteoform identification is c, the number of operations of the algorithm is proportional to $n^2 t^{c+1}$, where n is the number of nodes in the mass graph and t is the largest number of variable PTMs in a proteoform.

The size of a distant set $D(x_i, x_j, r)$ is $O(n^r l^r)$, where l is the largest number of edges between two nodes in G. In the implementation, each distance set is stored in a sorted list, and Steps 12 and 13 are performed by merging two sorted lists with $O(n^r l^r)$ steps. The time complexity of Steps 11-13 is $O(n^r l^{r+1})$. Similarly the number of operations of Steps 7-10 is also $O(n^r l^{r+1})$. The time complexity of Steps 5-13 is $\sum_{r=0}^{t} O(n^r l^{r+1}) = O(n^t l^{t+1})$, and the time complexity of the whole algorithm is $O(n^{t+2} l^{t+1})$.

The types of variable PTMs in proteoform identification are often limited. For example, only 5 types of PTMs were used in the experiments for the identification of proteoforms of the histone H4 protein. In this case, Algorithm 1 has a better time complexity. When a constant number c of PTM types are considered, the red edges in G can be divided into c types (variable PTMs). For example, the red edges in Figure 4.1(a) are divided into four types based on their corresponding PTMs: acetylation, methylation, phosphorylation, and dimethylation. Each path in G has a *modification vector* $[z_1, z_2, \ldots z_c]$ where z_i is the number of red edges corresponding to the ith type of PTM. For example, the modification vector of the bold path in Figure 4.1(a) is $[1, 1, 1, 0]$: one acetylation site, one methylation site, and one phosphorylation site. If two paths between two nodes have the same modification vector, they are consistent (their weights are the same) because their corresponding proteoforms have the same mass shifts introduced by PTMs. As a result, the size of a set $D(x_i, x_j, r)$ is bounded by the number of different modification vectors satisfying that $\sum_{i=1}^{c} z_i = r$, that is, the total number of red edges is r. The bound equals the number of ways to distribute r balls into c boxes, which is $O(r^c)$. Since the largest number of edges between two nodes $l \leq c+1$ is a constant, the time complexity of Steps 7-13 is $O(r^c)$. The number of operations in Steps 5-13 is $\sum_{r=0}^{t} O(r^c) = O(t^{c+1})$, and the time complexity of the whole algorithm is $O(n^2 t^{c+1})$.

Finding r-consistent preceding node pairs A node pair (u_1, u_2) in G and its r-distance set $(u_1, u_2, r) = \{d_1, d_2, \ldots, d_k\}$ are represented by triplets $< u_1, u_2, d_1 >, \ldots, < u_1, u_2, d_k >$. For a given r, the triplets of distance sets (u, v, r) for all node pairs (u, v) in G are merged and sorted based on the distance. Similarly, node pairs in H and their distances are also represented by a list of triplets sorted by the distance. The two sorted

triplet lists are compared to find the r-consistent preceding node pairs for all node pairs (u, v) satisfying that u is in G and v is in H. The number of operations in this step is proportional to $n^2 L \log(nL) + m^2 \log m + Z$, where L is the size of the largest r-distance set in G, m is the number of nodes in H, and Z is the total number of reported r-consistent node pairs.

Prefix residue masses in deconvoluted top-down MS/MS spectra may contain small errors introduced in measuring the m/z values of fragment ions. To address this problem, an error tolerance ϵ is used in finding r-consistent preceding node pairs. With the error tolerance, two paths are consistent if the difference of their weights is no larger than ϵ, and a triplet $< u_1, u_2, d_u >$ from G matches a triplet $< v_1, v_2, d_v >$ from H if $|d_u - d_v| \leq \epsilon$.

When the number of the types of variable PTMs in a proteoform is a constant, the algorithms for computing r-distance sets need polynomial time. In practice, we can further speed up the algorithms by removing some node pairs (u_1, u_2) from the computation. That is, we compute $D(u_1, u_2, r)$ only if the number of edges of the shortest path from u_1 to u_2 is no large than a user defined parameter L.

4.2.3 Algorithms for the RMGA problem

We present a dynamic programming algorithm (Figure 4.3) for the local RMGA problem. The algorithm can be modified to solve the semi-global and global RMGA problems. Let $x_0, x_1, \ldots, x_n$ be the nodes in the proteoform mass graph G in the topological order, and let $y_0, y_1, \ldots, y_m$ be the nodes in the spectral mass graph H in the topological order. We fill out a three dimensional table $T(i, j, k)$ for $0 \leq i \leq n$, $0 \leq j \leq m$, and $0 \leq k \leq t$. The value $T(i, j, k)$ is the highest shared mass counting score among all consistent path pairs (A, B) such that A ends at x_i and contains k red edges, and B ends at y_j. Let $C(i, j, r)$ be the set of all r-consistent preceding node pairs of (x_i, y_j). The values of $T(i, j, k)$ are computed using the following function:

$$T(i,j,k) = \begin{cases} \max\limits_{0 \le r \le k} \max\limits_{(x_{i'}, y_{j'}) \in C(i,j,r)} T(i', j', k-r) + 1 & \text{if } \cup_{r=0}^{k} C(i,j,r) \neq \emptyset; \\ 1 & \text{if } \cup_{r=0}^{k} C(i,j,r) = \emptyset \text{ and } j = k = 0; \\ -\infty & \text{otherwise.} \end{cases} \tag{4.1}$$

When (x_i, y_j) has no consistent preceding node pairs and $k = 0$, the value $T(i, j, 0)$ is set as 1 because two empty paths have a shared prefix weight 0. After all values in the table $T(i, j, k)$ are filled out, we find the largest one in the table and use backtracking to reconstruct a best scoring local alignment. The number of operations of the algorithm is proportional to t^2nmM, where M the size of the largest set $C(i, j, r)$.

The recurrence relation can be slightly modified to solve the semi-global and global RMGA problems. For the semi-global alignment problem, we change the second line in Equation (4.1) to $T(i, j, k) = 1$ if $\cup_{r=0}^{k} C(i, j, r) = \emptyset$ and $j = k = 0$, that is, y_j is required to be the start node. For the global alignment problem, we change the second line in Equation (4.1) to $T(i, j, k) = 1$ if $\cup_{r=0}^{k} C(i, j, r) = \emptyset$ and $i = j = k = 0$, that is, both x_i and y_j are required to be the start nodes.

4.3 Results

We developed TopMG (TOP-down mass spectrometry-based proteoform identification using Mass Graphs) based on the proposed algorithms using C++. All the experiments were performed on a desktop with an Intel Core i7-3770 Quad-Core 3.4 GHz CPU and 16 GB memory.

4.3.1 Evaluation on speed, memory usage, and accuracy

A test data set of PrSMs with mutations, which were treated variable PTMs, was generated from the EC data set for evaluating the speed, memory usage, and accuracy of TopMG. The proteome database of *Escherichia coli* K-12 MG1655 was downloaded from the UniProt database [72] (version June 18, 2015, 4 305 entries) and concatenated with a shuffled decoy database of the same size. All the 4 054 top-down MS/MS spectra from the EC data set

The algorithm for the local RMGA problem

Input: A proteoform mass graph G with nodes $x_0, x_1, \ldots, x_n$ in the topological order, and a spectral mass graph H with nodes $y_0, y_1, \ldots, y_m$ in the topological order, and a number t.

Output: A path A in G and a path B in H such that the number of red edges in A is no more than t and Score(A, B) is maximized.

1. **For** $i = 0$ to n **do**
2. **For** $j = 0$ to m **do**
3. **For** $k = 0$ to t **do**
4. **If** $k = 0$ **then** set $T(i, j, 0) = 1$ **else** set $T(i, j, k) = -\infty$.
5. **For** $r = 0$ to k **do**
6. **For** each node pair $(x_{i'}, y_{j'}) \in C(i, j, k - r)$ **do**
7. **If** $T(i', j', k - r) + 1 > T(i, j, k)$ **then** update $T(i, j, k) = T(i', j', k - r) + 1$.
8. Find the largest value of $T(i, j, k)$ for $0 \leq i \leq n$, $0 \leq j \leq m$, $0 \leq k \leq t$ and use backtracking to find a best scoring local alignment.

Figure 4.3: The algorithm for the local RMGA problem.

were deconvoluted by TopFD and then searched against the target-decoy concatenated EC proteome database using TopPIC [36]. In the database search, the error tolerances for precursor and fragment masses were set as 15 ppm and no mass shifts were allowed. A total of 861 PrSMs were identified with a 1% spectrum-level false discovery rate (FDR), which were further filtered by the number of matched fragment ions, resulting in 767 PrSMs with at least 15 matched fragment ions.

The 767 PrSMs without PTMs were used to generate test PrSMs with PTMs (mutations). Three mutations: lysine (K) to cysteine (C), threonine (T) to alanine (A), and valine (V) to glycine (G), were treated as variable PTMs. Let (P, S) be a PrSM between a spectrum S and a protein sequence $P = a_1 a_2 \ldots a_n$ without PTMs and truncations, and Ω a set of variable PTMs (mutations). We change the protein sequence P to introduce variable PTMs (mutations) into the PrSM. We first randomly select a mutation from amino acid x to y in Ω and an amino acid $a_i = y$ in P, then replace a_i with the amino acid x, resulting in a protein sequence P_1 with a mutation. In addition, a random amino acid sequence with a random length between 1 and 20 is appended to the N terminus of P_1, and another random

sequence with a random length between 1 and 20 is appended to the C-terminus of P_1. The PrSM between the resulting sequence and S contains a variable PTM (mutation), an N-terminal truncation, and a C-terminal truncation. Using this method, a total of 11 505 test PrSMs (15 for each of the 767 PrSMs) were generated. In addition, PrSMs with $2, 3, \ldots, 10$ PTMs and N- and C- terminal truncations were generated using a similar method. A total of 115 050 PrSMs were generated.

The semi-global mass graph alignment algorithm in TopMG was employed for identifying a top proteoform for each test PrSM. If the proteoform reported by TopMG has more than 15 matched fragment ions, we say TopMG identifies a PrSM. A reported proteoform may contain some mass shifts that are localized to several candidate PTM sites, not single ones. If one candidate site of a mass shift is correct, we say the mass shift is consistent with the correct site in the target proteoform. If a reported proteoform has the same N-terminal and C-terminal truncations as the target one and each mass shift in the reported proteoform is consistent with its corresponding PTM site in the target proteoform, the identification is correct.

We tested the running time, memory usage, accuracy of TopMG on the 11 505 test PrSMs with 5 variable PTMs each using various settings for L: $10, 20, 30, 40, 50, 60,\ 70, 80, 90, 100$ (see Section 4.2.2). The error tolerance ϵ was set as 0.1 Dalton (Da); the largest number of red edges (PTMs) t was set as 10; the three mutations were treated as variable PTMs. When the setting of L increases from 10 to 100, the running time increases from 328 minutes to 947 minutes, the memory usage increases from 1.2 GB to 2.2 GB, and the percentage of correctly identified proteoforms increases from 38.8% to 81.8% (Figure 4.4). TopMG achieved a good balance between the speed and the accuracy rate when $L = 40$. Of the 11505 test PrSMs, TopMG ($L = 100$) reported 11308 (98.3%) PrSMs with at least 15 matched fragment ions, 11101 (96.5%) PrSMs with correct N- and C-terminal truncations, and 11019 (95.7%) PrSMs with both correct terminal truncations and correct numbers of variable PTMs. Most incorrectly identified proteoforms contained some PTMs that were not correctly localized because of the existence of random matches between experimental fragment masses and theoretical prefix residue masses.

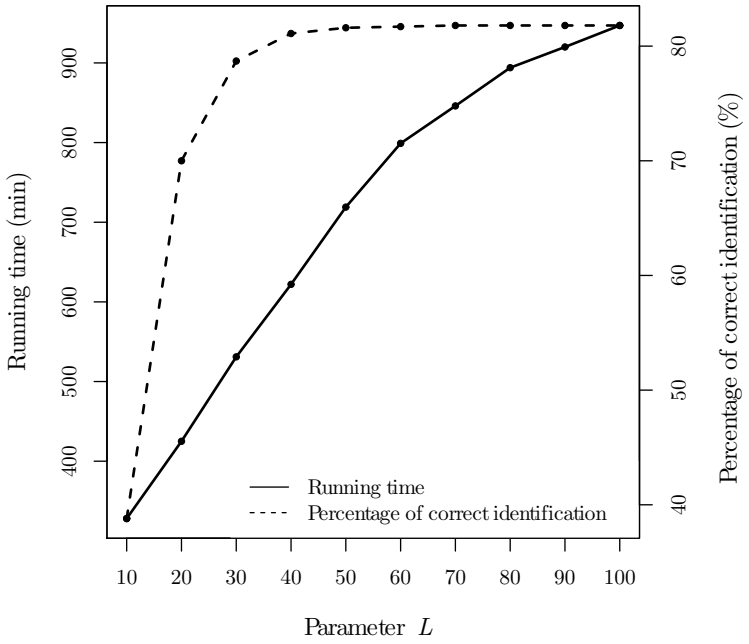

Figure 4.4: The running time and percentages of correctly identified PrSMs for the 11505 test PrSMs with 5 variable PTMs each when the parameter L is set as $10, 20, \ldots, 100$.

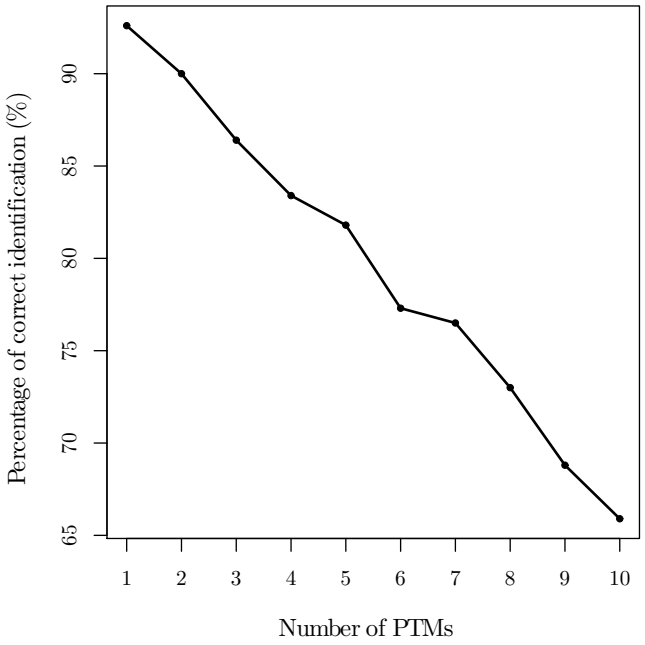

Figure 4.5: The percentages of correctly identified PrSMs for the test PrSMs with various numbers of variable PTMs.

We tested the accuracy rates of TopMG on the test PrSMs with various numbers (1 to 10) of variable PTMs, in which the parameter L was set as 40 and all other parameters were set as the same as the previous experiment. When the number of variable PTMs increases from 1 to 10, the accuracy rate decreases from 92.6% to 65.9% (Figure 4.5). Of the 11505 test PrSMs with 10 variable PTMs, TopMG reported 11019 (95.7%) PrSMs with at least 15 matched fragment ions, 10552 (91.7%) PrSMs with correct N- and C-terminal truncations, and 10056 (87.4%) PrSMs with both correct terminal truncations and correct numbers of variable PTMs, showing that most of the incorrectly identified proteoforms contained incorrectly localized PTMs.

4.3.2 Proteoform identifications from the histone data sets

We deconvoluted all the MS/MS spectra in the histone data sets using TopFD. Five common variable PTMs in the histone protein (Table 3.1) were included in the construction of

proteoform mass graphs. For precursor masses, ± 1 and ± 2 Da errors were allowed, which may be introduced by the deconvolution algorithm. For a spectrum with a precursor mass m, we generated five candidate spectra with precursor masses $m-2$, $m-1$, m, $m+1$, $m+2$, respectively, and the spectrum with the best alignment result was reported. The error tolerance ϵ was set as 0.1 Da and the largest number of red edges t was set as 10; the parameter L was set as 40.

By aligning the spectra against the proteoform mass graph, TopMG (the algorithm for the semi-global RMGA problem) identified from the first histone data set 1087 PrSMs with at least 10 matched fragment ions, including 918 matches with at least 20 matched fragment ions (Figure 4.6(a)). Of the 1087 matches, 239 contain more than 3 PTM sites (Figure 4.6(b)).

The running time of TopMG was about 88 minutes. The running time depends on the sizes of the r-distance sets and the numbers of r-consistent preceding node pairs reported from the proteoform and spectral mass graphs. For the histone H4 protein with the five variable PTMs, the size of the largest r-distant set was 553. For each spectral mass graph, we count the total number N of the consistent preceding node pairs used in the mass graph alignment algorithm, that is, $N = \sum_i \sum_j \sum_{r=0}^{t} C(i,j,r)$. The average value of N for all the $3,252$ spectra was 5.60×10^6, and the maximum value of N was 6.20×10^7.

We compared the performance of TopMG and MS-Align-E [35] on the first histone data set. For MS-Align-E, the error tolerance for fragment masses was set as 15 ppm and all the other parameters were set as the same as TopMG. The running time of MS-Align-E was 505 minutes. MS-Align-E identified 1 031 PrSMs with at least 10 matched fragment ions. TopMG identified 991 of 1 031 matches reported by MS-Align-E as well as 96 PrSMs missed by MS-Align-E, all of which correspond to proteoforms with terminal truncations. The main reason why 96 PrSMs were missed by MS-Align-E is that MS-Align-E is not able to identify truncated proteoforms. The comparison demonstrated that TopMG outperformed MS-Align-E in identifying truncated proteoforms. TopMG missed 40 PrSMs identified by MS-Align-E because it may fail to identify PrSMs with very low sequence coverage with the parameter setting $L = 40$. When L was set as 200, TopMG identified all the 40 PrSMs.

Proteoforms reported by TopMG tend to have more matched fragment ions (Figure 4.6(a)) and less PTM sites (Figure 4.6(b)) compared with those reported MS-Align-E.

The second histone data set contains 1 349 CID and 1 349 ETD spectra of the histone H4 protein. TopMG identified from these spectra 1 051 PrSMs of the histone H4 protein with at least 10 matched fragment ions, including 851 matches with at least 20 matched fragment ions. Of the 1 051 matches, 291 contain more than 3 PTM sites. Coupled with the Thrash algorithm [19], the absolute mass mode of ProSightPC reported 89 proteoforms as well as their corresponding PrSMs with at least 10 matched fragment ions from these spectra. TopMG identified all the 89 spectra corresponding to the 89 matches reported by ProSightPC. In addition, TopMG identified 79 PrSMs whose precursor masses cannot match any proteoforms reported by ProSightPC, showing that the corresponding proteoforms are missed by ProSightPC. Manual inspection confirmed that a proteoform with an N-terminal truncation (18 amino acids are removed) was identified by TopMG, but missed by ProSightPC. TopMG also identified proteoforms missed by ProSightPC from the spectra of the histone H2A, H2B, and H3 proteins in the second histone data set.

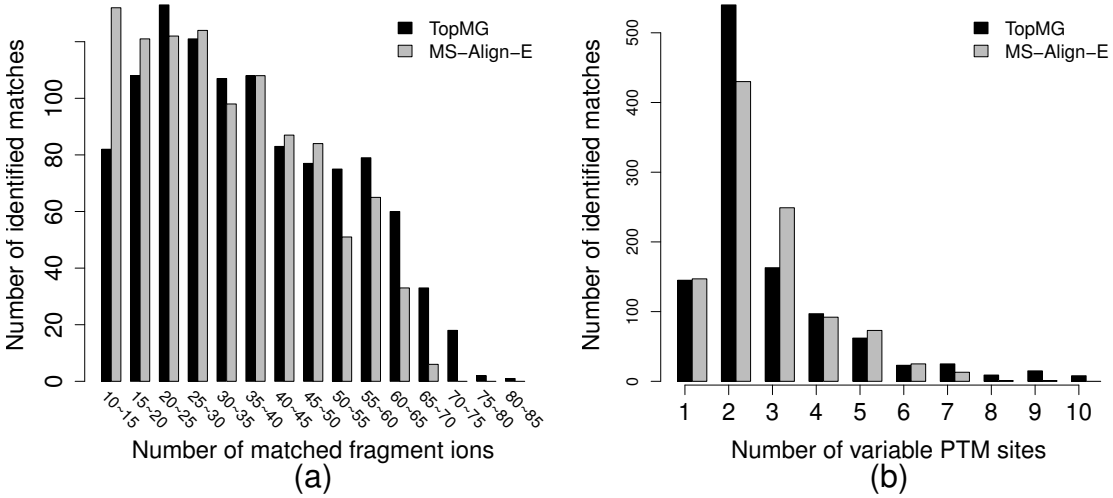

Figure 4.6: Histograms for the PrSMs reported from the first histone data set by TopMG with $L = 40$ and MS-Align-E: (a) the number of matched fragment ions; (b) the number of variable PTM sites.

4.4 Discussion

Unlike splicing graphs [75] and variant graphs [78], amino acid residue masses are stored as weights of edges, not of nodes, in mass graphs. Suppose residue masses are stored as weights of nodes. Let u_1, u_2, u_3 be the three nodes representing the first arginine (R) and its modified forms R[Acetylation] and R[Methylation] in the protein in Figure 4.1 and v_1, v_2, v_3, v_4 be the four nodes representing the first lysine (K) and its modified forms K[Aceyltation], K[Methylation] and K[Dimethylation]. We need 12 edges to connect all node pairs (u_i, v_j) for $1 \leq i \leq 3$ and $1 \leq j \leq 4$, making the graph more complex than the mass graph representation. The example shows that using edge weights in graphs is more efficient than node weights in representing proteoforms with variable PTMs.

The mass graph alignment problem is similar to the spliced alignment problem [83], but they are different. The spliced alignment problem studies sequence alignment, not mass alignment. In a sequence alignment problem, a substitution in a sequence does not significantly affect the alignment results. For example, changing "A" to "T" in x in the sequence alignment between $x = ACGT$ and $y = ACGT$ does not affect the matching pairs of CGT. However, this property does not hold for mass alignment. For example, the red bold path in Figure 4.1(a) and the path from y_0 to y_6 in Figure 4.1(b) has a shared mass counting score 4 because they share 4 prefix masses 0, 198, 340, and 521. If we change the mass on the red edge between x_1 and x_2 from 198 to 156, the two paths share only one prefix residue mass 0. The reason is that the mass shift "propagates" to all non-zero prefix residue masses of the red bold path. The "propagation" property makes mass alignment more challenging than sequence alignment.

Compared with MS-Align-E [35] and pTop [37], the main advantage of TopMG is that it is capable of identifying proteoforms with terminal truncations. Although using MS-Align-E or pTop to search spectra against a database containing all possible proteoforms with terminal truncations can also identify truncated proteoforms, the size of the database is extremely large, making the approach inefficient. For example, a protein sequence with 300 amino acids has 45 150 different truncated forms.

The parameter L determines the sensitivity and speed of TopMG. The experiments showed setting $L = 40$ obtained a good balance between speed and sensitivity. In practice, users can adjust the setting of L to satisfy specific requirements in data analyses. When a long running time is acceptable, the setting of L can be increased to 100 or even the length of the target protein to increase the sensitivity of TopMG.

CHAPTER 5

STATISTICAL SIGNIFICANCE ESTIMATION FOR IDENTIFIED COMPLEX PROTEOFORMS

5.1 Introduction

Assigning accurate statistical significance to proteoform identifications is an important step in top-down mass spectral interpretation [74, 84]. In spectral identification, a query spectrum is searched against a protein sequence database to find several candidate proteoform spectrum matches (PrSMs). These matches are usually ranked by their E-values to find the best one. In proteome-level MS studies, thousands of spectra are searched and matched to proteoforms, and these identified PrSMs are often filtered by an E-value cutoff. Accurate E-values of identifications efficiently distinguish correct identifications from incorrect ones and increase the number of identifications.

Many efforts have been made to develop methods for estimating the statistical significance of identifications in bottom-up MS [85], in which proteins are digested into short peptides before MS analysis. Because of the similarity between bottom-up MS and top-down MS, most of the methods developed for bottom-up MS can be used in top-down MS.

There are three types of methods for assigning statistical significance to identifications in bottom-up MS. The first is *probability distribution fitting*, which has been widely used [86–89]. In this approach, a parametric probability distribution is fit to an empirical score distribution and then used to compute the statistical significance of identifications. Methods using probability distribution fitting highly depend on the empirical score function in spectral identification and may fail to to accurately compute extremely small p-values or E-values [90].

The second is the *generating function* method [73, 90], which provides an analytical framework for assigning statistical significance to identifications. Given a match between a query spectrum and a peptide with a score t, its p-value is computed as follows: a dynamic programming algorithm is employed to compute the distribution of the similarity score

between the spectrum and a random peptide whose molecular mass matches the precursor mass of the spectrum, and then the p-value is computed based on the probability that the score is no less than t in the distribution. This approach is capable of accurately assigning p-values to identifications. When thousands of spectra are analyzed, the score distribution of each query spectrum needs to be computed separately, making it much slower than the first approach.

The third is the *Markov chain Monte Carlo* (MCMC) method [91]. Importance sampling methods, such as direct probability distribution (DPR), are often used in Monte Carlo simulation to estimate probabilities of extremely rare events [92]. Mohimani et al. [93] proposed MS-DPR, which successfully applied MCMC with DPR to estimate the statistical significance of identified cyclic peptides. In MS-DPR, peptides are sampled by a random walk on a Markov chain to estimate the distribution of the similarity score between a query spectrum and a random peptide as well as the p-value of an identification.

Many proteoform identifications in top-down MS contain multiple alterations, especially multiple variable PTMs [4, 94]. The problem of assigning statistical significance to identifications with multiple PTMs has not been extensively studied. In bottom-up MS, peptide identifications seldom contain three or more PTMs, and there is no urgent need to solve the problem. In top-down MS, most existing methods are extended from those in bottom-up MS, which are not designed for the problem.

When variable PTMs are allowed, many proteoforms of a protein are similar, and the similarity scores of a query spectrum and these proteoforms are not independent. As a result, it is a challenging problem to accurately estimate proteoform-level statistical significance of identifications. In this chapter, we focus on the estimation of protein-level statistical confidence of identifications.

The first two approaches in bottom-up MS have been applied to estimate the protein-level statistical significance of identifications in top-down MS. In ProSightPC [30], the distribution of similarity scores of proteoform identifications is fit to a Poisson distribution for p-value estimation. The generating function method was extended to handle unexpected alterations in proteoform identifications [74] and used in MS-Align+ [32], TopPIC [36], and MS-PathFinder [18].

In this chapter, we propose TopMCMC, an MCMC method with DPR for estimating the protein-level statistical significance of proteoform identifications with multiple PTMs identified by top-down MS. Because of the existence of PTMs, the MS-DPR method proposed by Mahimani et al. [93] cannot be directly applied to solve this problem. We designed a new Markov chain model for representing proteins in top-down spectral identification and a fast greedy algorithm for computing the similarity score between a query spectrum and a protein with variable PTMs. By combining the Markov chain and the greedy algorithm, TopMCMC is capable of efficiently assigning protein-level statistical significance to PrSMs. We used two methods to evaluate the performance of TopMCMC on four top-down MS data sets, and showed that TopMCMC achieved high accuracy in estimating p-values of identifications. By coupling TopMCMC and spectral alignment algorithms in TopMG [1], we identified more top-down mass spectra from an MCF-7 data set than TopMG with the generating function method.

5.2 Methods

5.2.1 Similarity scores of PrSMs

In proteoform identification, a score is reported for each identified PrSM to evaluate the similarity of the match, and the statistical significance of the match is estimated based on the similarity score. Next we describe the representations of spectra and proteins, and define a similarity score between an MS/MS spectrum and a proteoform.

In preprocessing of top-down mass spectra, spectral deconvolution tools [18,19,23] are often used to convert complex tandem mass spectra to neutral monoisotopic fragment masses. A deconvoluted tandem mass spectrum S is represented by a monoisotopic precursor mass and a list of neutral monoisotopic fragment masses. The residue mass of S is defined as $\text{PrecMass}(S) - \text{mass}(\text{H}_2\text{O})$, where $\text{PrecMass}(S)$ is the monoisotopic neutral precursor mass of S and $\text{mass}(\text{H}_2\text{O})$ is the monoisotopic mass of a water molecule.

A proteoform F of n amino acids (some amino acids may be modified) is represented as a list of n integer residue masses, that is, $F = a_1 a_2 \dots a_n$, where a_i is the integer residue mass of the ith amino acid. In practice, residue masses of amino acids are discretized by

multiplying them by a scale factor and rounding the results to integers [35]. The residue mass of the protein P is the sum of its amino acid residue masses, $\text{mass}(F) = \sum_{i=1}^{n} a_i$.

To compute the similarity between spectrum S and proteoform F, we generate a theoretical fragment mass list of F. For $1 \leq i \leq n-1$, the mass $f_i = \sum_{k=1}^{i} a_k$ is called a prefix residue mass of F; the mass $g_i = \sum_{k=n-i+1}^{n} a_k$ is called a suffix residue mass of F. Combining all the prefix and suffix residue masses gives us a theoretical mass list of F, denoted by $t(F) = \{f_1, \ldots, f_{n-1}, g_i, \ldots, g_{n-1}\}$. The theoretical mass list contains neutral monoisotopic fragment masses of b- and y-ions, which are used in the interpretation of CID spectra. We add mass shifts to prefix and suffix residue masses to generate theoretical mass lists for other dissociation methods. For example, when the scale factor in discretization is 1, a mass shift of 17 is added to all prefix residue masses to obtain theoretical c-ion masses, which are commonly observed in ETD spectra.

The mass counting score between S and F is defined based on their residue masses and matched fragment masses. If the residue mass of S matches the residue mass of F, the mass counting score $\text{FScore}(S, F)$ is defined as the number of matched fragment masses between S and $t(F)$. Otherwise, the similarity score is 0. The mass counting score is used as the similarity score of a spectrum and a proteoform in the following analysis.

5.2.2 Similarity scores between proteins and spectra

Database search is the most used method for proteoform identification by top-down MS. Many protein databases contain only unmodified protein sequences, not proteoforms with modifications. Several variable PTMs are often provided by the user to identify modified proteoforms.

Let V be a multiset of variable PTMs. Each PTM in V is represented by its discretized monoisotopic mass shift. Similar to residue masses, mass shifts of PTMs are discretized by multiplying them by a scale factor and rounding the results to integers. To simplify the analysis, we assume that a PTM $v \in V$ can modify any amino acid with a residue mass a if the modified residue mass is positive, that is, $a + v > 0$. In Section 5.2.6, we will discuss the case in which a PTM modifies only one or several amino acids. A PTM may occur several times in the multiset V. For example, $V = \{80, 80, 16\}$ specifies two phosphorylation sites

and one oxidation site in a modified proteoform. A length n proteoform F is a modified proteoform of a length n protein P with PTMs $V = \{v_1, v_2, \ldots, v_k\}$ if (1) there are $n - k$ matched mass pairs in P and F and (2) the multiset of the mass differences of the remaining k mass pairs is the same as V. For example, $57, \mathbf{147}, 114, 156, 129, \mathbf{167}, 128$ is a modified proteoform of protein $57, \mathbf{131}, 114, 156, 129, \mathbf{87}, 128$ with two PTMs $V = \{16, 80\}$.

Let $D(P, V)$ be the set of all modified proteoforms of a protein P with a multiset $V = \{v_1, v_2, \ldots, v_k\}$ of PTMs. The P-score between S and P with the multiset V is the maximum similarity score between S and the proteoforms in $D(P, V)$, denoted by $\text{PScore}(S, P, V)$. That is, $\text{PScore}(S, P, V) = \max_{F \in D(P,V)} \text{FScore}(S, F)$. All proteoforms in $D(P, V)$ have the same residue mass $m + \sum_{i=1}^{k} v_i$, where m is the residue mass of P. When $m + \sum_{i=1}^{k} v_i$ does not match the residue mass of S, the score $\text{PScore}(S, P, V)$ is zero.

In this chapter, we study protein-level statistical significance of matches between proteins and spectra. When $\text{PScore}(S, P, V) = t > 0$, we use an MCMC-based method to estimate the probability that the P-Score between the spectrum and a random protein with n amino acids and a residue mass m is no less than t.

It is inefficient to compute $\text{PScore}(S, P, V)$ by enumerating all proteoforms in $D(P, V)$. The size of $D(P, V)$ is proportional to n^k, where n is the length of P and k is the size of V. When the PTM list V is long, the size $D(P, V)$ is very large.

Spectral alignment can be used for computing the P-score between a spectrum and a protein. The dynamic programming method for spectral alignment solves the combinatorial explosion problem by filling out a 3-dimensional table. Although it is fast for aligning a spectrum-protein pair, it is still inefficient for the MCMC method, in which tens of thousands of random proteins need to be aligned with a query spectrum.

Here we use a greedy algorithm (Figure 5.1) to quickly obtain an estimation of $\text{PScore}(S, P, V)$. Two proteoforms F_1 and F_2 in $D(P, V)$ are neighbors if we can obtain F_2 from F_1 by shifting the position of one PTM in F_1 and *vice versa*. For example, $F = 57, \mathbf{147}, 114, 156, 129, \mathbf{167}, 128$ is a proteoform of protein $57, 131, 114, 156, 129, 87, 128$ with two PTMs $\{16, 80\}$, and $F' = 57, 131, 114, \mathbf{172}, 129, \mathbf{167}, 128$ is a neighbor proteoform of F. The proteoform F' can be obtained from F by shifting the position of the PTM 16 to the right: the PTM is shifted from the second amino acid residue to the fourth. In the greedy

algorithm, we start with a random proteoform F in $D(P, V)$. In each round, we select a proteoform F' from all neighbors of F to maximize the score $\text{FScore}(S, F')$ and use F' to replace F. The algorithm is terminated if the similarity score cannot be improved, and the final score is used as an estimation of $\text{PScore}(S, P, V)$.

A greedy algorithm for estimating similarity scores

Input: A protein sequence P, a spectrum S, and a multiset V of PTMs.

Output: An estimation of the similarity score $\text{PScore}(P, S, V)$.

1. Randomly select a proteoform F in $D(P, V)$.
2. **Repeat**
3. Find a proteoform F' in all neighbors of F such that $\text{FScore}(S, F')$ is maximized.
4. Set the score difference $\delta \leftarrow \text{FScore}(S, F') - \text{FScore}(S, F)$.
4. **If** $\delta > 0$ **then** $F \leftarrow F'$
5. **Until** $\delta \leq 0$
6. Report the score $\text{FScore}(S, F)$.

Figure 5.1: A greedy algorithm for estimating similarity scores.

5.2.3 Markov chains representing proteins

Similar to the method proposed by Mohimani et al. [93], we assume that the alphabet of protein sequences is not the masses of the 20 standard amino acids, but the set of all positive integers $Z^+ = \{1, 2, \ldots\}$. Using the alphabet of Z^+ makes it possible to build a homogeneous Markov chain for representing all proteins that match a query spectrum.

Let $\Omega_{n,m}$ be the collection of all length n proteins with a residue mass m, in which the probabilities of the elements follow a uniform distribution. Next we define sister proteins and introduce a method for building a Markov chain representing $\Omega_{n,m}$.

Two masses a_i and b_i $(1 \leq i \leq n)$ in two proteins $a_1 a_2 \ldots a_n$ and $b_1 b_2 \ldots b_n$ are a matched mass pair if $a_i = b_i$, and a mismatched mass pair otherwise. Two proteins are *sister proteins* if they have the same length and the same residue mass, and contain at most 2 mismatched mass pairs. For example, $57, \mathbf{71}, 114, 156, \mathbf{129}, 57, 128$ and $57, \mathbf{87}, 114, 156, \mathbf{113}, 57, 128$ are sister proteins. They have the same length 6, the same residue mass 712, and contain only

two mismatched mass pairs $(71, 87)$ and $(129, 113)$, whose mass differences are opposites: 16 and -16. In addition, a protein is a sister protein of itself by definition.

Below we give the total number of sisters of a protein $P = a_1 a_2 \dots a_n$ with a residue mass $m = \sum_{i=1}^{n} a_i$. Let $P' = b_1 b_2 \dots b_n$ be a sister protein of P with two mismatched mass pairs: (a_i, b_i) with $a_i > b_i$ and (a_j, b_j) with $a_j < b_j$. There are a total of $a_i - 1$ possible values for b_i, so the total number of such sister proteins is $(a_i - 1)$. For a given pair (a_i, b_i), there are $n-1$ possible positions for the other pair (a_j, b_j). As a result, the total number of sister proteins of P with two mismatched mass pairs is $\sum_{i=1}^{n}(a_i - 1)(n-1) = (m-n)(n-1)$. In addition, P is a sister protein of itself. The total number of sister proteins of P is $(m-n)(n-1)+1$.

We build a Markov chain C for the sample space $\Omega_{n,m}$ as follows. Each protein in $\Omega_{n,m}$ is represented by a state in C, and a state is connected to another state by a directed edge if and only if their corresponding proteins are sisters (Figure 5.2). Each state has an outdegree of $(m-n)(n-1)+1$ because its corresponding protein has $(m-n)(n-1)+1$ sister proteins. The transition probability of each edge is $\frac{1}{(m-n)(n-1)+1}$. The Markov chain is ergodic and aperiodic because it is connected and contains length-1 cycles. Based on the fundamental theorem of Markov chains [95], the Markov chain has a unique stationary distribution. In addition, the Markov chain C is homogeneous because each state in C has the same number of edges connecting to it and the transition probability for each edge is the same. It can be proved that the stationary distribution of C is a uniform distribution: each state has the same probability $\frac{1}{|\Omega_{n,m}|}$, where $|\Omega_{n,m}|$ is the size of the set $\Omega_{n,m}$. We will use the MCMC method to sample elements in $\Omega_{n,m}$.

5.2.4 The direct probability redistribution method

Let X be a random variable for the similarity score $\text{PScore}(S, P, V)$ between a spectrum S and a random protein $P \in \Omega_{n,m}$ with a fixed multiset $V = \{v_1, v_2, \dots, v_k\}$ of PTMs. The space of X is $\{0, 1, \dots, m\}$, where m is the number of masses in the spectrum S. When the spectrum S and multiset V are fixed, the score $\text{PScore}(S, P, V)$ is also defined as the score of the state in the Markov chain C corresponding to P. We use the MCMC random walk method to generate random proteins for estimating the distribution of X.

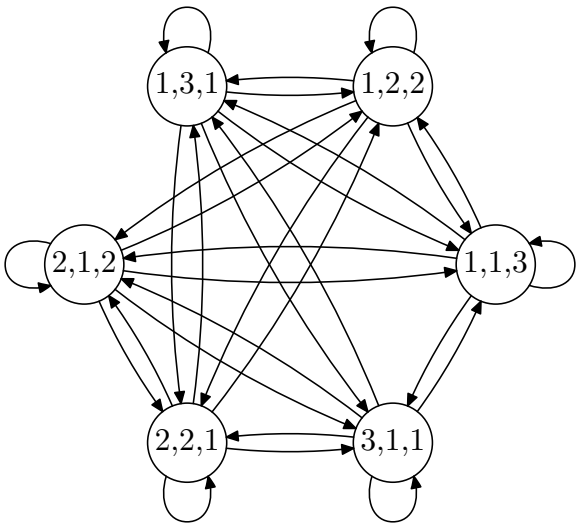

Figure 5.2: An example Markov chain for the sample space $\Omega_{3,5}$, which contains all proteins with length 3 and residue mass 5. Each protein is represented as a state in the Markov chain, and a state is connected to another if and only if their corresponding proteins are sister proteins. There are no edges connecting $(1,3,1)$ and $(2,1,2)$ because they contain 3 mismatched mass pairs. Each state is connected to itself because each protein is a sister protein of itself. Each state has an outdegree of $(m-n)(n-1)+1 = (5-3)(3-1)+1 = 5$. The transition probability of each edge is $\frac{1}{5}$.

In MS-DPR, two mismatched mass pairs in two sister peptides need to be neighbors, but those in two sister proteins in TopMCMC may be not neighbors. The definition of sister proteins in TopMCMC leads to abrupt changes of similarity scores of states visited in random walks and makes it possible to move from a state with a low score to another state with a high score with several transitions.

For an identified PrSM with a similarity score t, we need to estimate the probability $\Pr(X \geq t)$ to obtain its p-value. The probability is often very small when the score t is large. For example, the probability is usually less than 10^{-10} when $t = 20$. In the MCMC random walk method, billions of simulations (trial runs) are required to accurately estimate such a small probability. To speed up the computation, we need to oversample rare events to reduce the number of simulations.

The *Direct Probability Redistribution* (DPR) method is an efficient technique for reducing the number of simulations in estimating rare event probabilities in Monte Carlo simulation [92]. Let p_i $(0 \leq i \leq m)$ be the probability that $X = i$. The DPR method increases the transition probability of the edge from a state Q_1 to another state Q_2 if the

score for Q_2 is higher than that for Q_2. The oversampling procedure is a recursive function (Figure 5.3). Let $u_0 \leq u_1 \leq \ldots \leq u_m$ be oversampling factors, where u_i is the oversampling factor for states with score i. We assume that the oversampling factor increases when the score increases. In each iteration of the algorithm, a new state Q' is randomly selected from a current state Q using the Markov chain. The number of simulations starting from the new state Q' is based on its score s', the score s for Q, and a threshold $h \leq s$. There are three cases: (1) If the score s' is smaller than the threshold h, the number of simulations from state Q' is reduced to 0 (Step 4). (2) If the score s' is larger than s, the number of simulations from Q' is increased (Steps 6-8). (3) If the score s' is between h and s, that is, $h \leq s' \leq s$, the number of simulations from Q' is 1 (Step 9). The output of the procedure is stored in a list of counts $z_0, z_1, \ldots, z_m$, in which z_i represents the number of visited states with a score i. For each score i ($0 \leq i \leq m$), the stationary probability p_i is computed as $\frac{z_i/u_i}{\sum_{k=0}^{m} z_k/u_k}$. More details of the DPR method can be found in [92].

The oversampling factors $u_0, \ldots, u_m$ are important parameters for accurate estimation of rare event probabilities. Haraszti et al. [92] proved that $u_i = 1/p_i$ are the optimal oversampling factors. Since the stationary probabilities $p_0, \ldots, p_m$ are unknown, an iterative method is used to find settings for the oversampling factors (Figure 5.4). In the first iteration, the oversampling factors are set to $u_0 = \cdots = u_m = 1$ to estimate $p_0, \ldots, p_m$; in the next iterations the oversampling factors are set to $u_0 = 1/p_0, \ldots, u_m = 1/p_m$. The algorithm will be terminated after T iterations. The parameter T was set to 3 in the experiments.

5.2.5 Expected values of PrSMs

Given a spectrum S, a multiple set V of PTMs, and a random protein sequence P from $\Omega_{i,j}$, the DPR method is used to estimate the distribution of $\text{PScore}(S, P, V)$ when the sum of the residue mass j and the masses in V equals the residue mass of S. Let D be a protein sequence database that contains random sequences with various lengths and residue masses. We denote by $D_{i,j}$ the set of protein sequences in D with i amino acids and a residue mass j. The size of $D_{i,j}$ is denoted by $d_{i,j}$. In practice, the value $d_{i,j}$ is obtained by counting the number of protein sequences with i amino acids and residue mass j in the

MCMC simulation using DPR

Global variables: A Markov chain C, a total number c_{max} of simulations, a query spectrum S, a multiset V of PTMs, oversampling factors $u_0 \leq u_1 \leq \cdots \leq u_m$, and state counts $z_0, z_1, \ldots, z_m$ with initial values all set to 0.

Input: An initial state Q in the Markov chain C, a counter c for the number of sampled states, and a threshold h. The initial values for c and h are 0.

Output: The state counts $z_0, z_1, \ldots, z_m$.

1. **Procedure** Simulate(Q, c, h)
2. **While** $c < c_{max}$ **do**
3. Randomly select a next state Q' from Q using the Markov chain. The scores of Q and Q' are represented by s and s', respectively.
4. **If** $s' < h$ **then** return
5. **If** $s' > s$ **then**
6. **For** $i = 1$ **to** $\lfloor u_{s'}/u_s \rfloor - 1$ **do**
7. Randomly select x from $[u_s, u_{s'}]$ and find a score h' such that $u_{h'-1} \leq x \leq u_{h'}$.
8. Simulate(Q', c, h')
9. Set $c \leftarrow c + 1$, $Q \leftarrow Q'$ and $z_s \leftarrow z_s + 1$.

Figure 5.3: MCMC simulation using DPR.

protein sequence database used in top-down spectral identification. Each sequence in $D_{i,j}$ is randomly selected from the set $\Omega_{i,j}$. Let $X(i, j, t, V)$ be a random variable representing the number of protein sequences P in $D_{i,j}$ with $\text{PScore}(S, P, V) \geq t$. Note that $X(i, j, t, V)$ is zero when the sum of the residue mass j and the masses in V does not match the residue mass of S. The expected value of $X(i, j, t, V)$ is estimated to be $p(i, j, t, V) \cdot d_{i,j}$, where $p(i, j, t, V) = \Pr(\text{PScore}(S, P, V) \geq t)$. Let $X(t, V)$ be a random variable representing the number of proteins in D with a score $\text{PScore}(S, P, V) \geq t$. The expected value of $X(t, V)$ is $\sum_i \sum_j p(i, j, t, V) \cdot d_{i,j}$.

In top-down spectral identification, a set T of possible PTM types, instead of a multiset of PTM sites, is allowed in identified proteoforms. Let Φ_k be a set of all multisets V each containing *at most* k PTMs (may have repetitions) in T. We define a random variable $Y(k, t) = \sum_{V \in \Phi_k} X(t, V)$, which represents the number of pairs (P, V) with a score

Algorithm for estimating oversampling factors

Input: A Markov chain C, a query spectrum S, and a parameter T of iterations.

Output: Oversampling factors $u_0, u_1, \ldots, u_m$.

1. Set $u_0 = u_1 = \ldots = u_m = 1$.
2. **For** $i = 1$ to T **do**
3. Use the DPR method to estimate the state counts $z_0, z_1, \ldots z_m$ with the Markov chain C, the query spectrum S, and the oversampling factors.
4. For $i = 0, 1, \ldots, m$, compute $p_i = \frac{z_i/u_i}{\sum_{i=1}^{m} z_k/u_k}$.
5. Set $u_0 = 1/p_0, u_1 = 1/p_1, \ldots, u_m = 1/p_m$.
6. **Return** oversampling factors $u_0, u_1, \ldots, u_m$.

Figure 5.4: The algorithm for estimating oversampling factors.

$\text{PScore}(S, P, V) \geq t$, where P is a protein in D and V is a multiset in Φ_k. The expectation of $Y(k, t)$ is computed as $\sum_{V \in \Phi_k} \sum_i \sum_j p(i, j, t, V) \cdot d_{i,j}$. The expected value of $Y(k, t)$ is reported as the E-value for a PrSM with k variable PTM sites and a mass counting score t identified by database search. The p-value of the PrSM is the probability that the maximum score $\max_{P \in D, V \in \Phi_k} \text{PScore}(S, P, V) \geq t$, which equals the probability that at least one match between a protein P in D and a multiset $V \in \Phi_k$ has a score $\text{PScore}(S, P, V) \geq t$. That is, the p-value of the PrSM is the probability $\Pr(Y(t, V) \geq 1)$. Because it is complicated to compute the probability, we use a simple method to estimate it.

To speed up the computation, the greedy algorithm in Figure 5.1 is used to estimate P-scores in the DPR method. Below we describe how to estimate the probability $p(i, j, t, V)$ with the greedy algorithm. Consider an identified PrSM (S, P^*) between a spectrum S and a protein P^* with a multiset V of PTMs and a similarity score t. We first use the greedy algorithm to compute an estimation t' of $\text{PScore}(S, P^*, V)$. Second, we use the DPR method to compute the probability that the estimation of $\text{PScore}(S, P, V)$ reported by the greedy algorithm is no less than t', where P is a random protein in $\Omega_{i,j}$. The probability is used as an estimation of $p(i, j, t, V)$.

5.2.6 Sequences of standard amino acids

In the Markov chain model described previously, the alphabet of a protein sequence is all positive integer numbers, not the residue masses of the 20 amino acids. We modify the model to sample protein sequences of the 20 amino acids.

In the modified model, the alphabet contains 19 integer masses, each of which is the discretized residue mass of an amino acid. We use 19 instead of 20 masses because leucine and isoleucine have the same integer mass value and are treated as the same. Because of the small size of the alphabet, two sister proteins $a_1a_2 \ldots a_n$ and $b_1b_2 \ldots b_n$ of the 19 masses often have two mismatched mass pairs (a_i, b_i) and (a_j, b_j) where $a_i = b_j$ and $a_j = b_i$. That is, the two proteins have the same composition of amino acids. As a result, simulations in the MCMC method may be limited to sequences similar to that of the initial state.

To address the problem, we introduce cousin proteins, which allow more changes in sequences compared with sister proteins. The lengths of two cousin proteins can be different, and they have at most two pairs of mismatched segments, the length of which can be longer than one. A protein with two mismatched segments is divided into 5 segments by the four ending points of the two mismatched segments. Two protein sequences P_1 and P_2 are cousin proteins if they have the same residue mass and can be represented by concatenations of three matched segments and two mismatched segments $P_1 = A_1A_2A_3A_4A_5$ and $P_2 = B_1B_2B_3B_4B_5$, where $A_1 = B_1$, $A_2 \neq B_2$, $A_3 = B_3$, $A_4 \neq B_4$, and $A_5 = B_5$. That the segments $A_1, A_3, A_5, B_1, B_3, B_5$ may be empty ones. In addition, a protein sequence is a cousin of itself.

Because cousin proteins may have various lengths, a Markov chain in the modified model represents protein sequences with various lengths, not a fixed length. Let Ω_j be the set of all protein sequences on the alphabet of the 19 masses with a residue mass j. Each state in the Markov chain represents a protein in Ω_j. Two states are connected by an edge if their corresponding proteins are cousins. In the implementation of the method, we added an additional constraint to reduce the number of cousin proteins of a state: the lengths of A_2 and A_4 are each no longer than 2. In addition, an error tolerance is allowed for the residue masses of two cousin proteins. An example of cousin proteins is given in Figure 5.5.

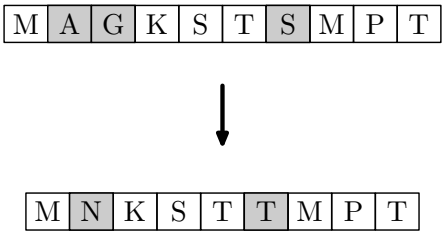

Figure 5.5: An example of cousin proteins on the alphabet of the residue masses of the 20 standard amino acids. The sum of the residue masses in the substrings 'AG' and 'S' in the protein MAGKSTSMPT is the same as that in the substrings 'N' and 'T' in the protein MNKSTTMPT within an error tolerance of 15 ppm.

The number of cousins of a random protein in Ω_j is not fixed because the proteins in Ω_j have various lengths and the numbers of possible mismatch segment pairs (A_2, B_2) and (A_4, B_4) of proteins are not fixed. As a result, we assign different transition probabilities to edges. For a state corresponding to a protein with k cousin proteins, we assign a transition probability $\frac{1}{k}$ to each edge leaving the state. The stationary distribution of such a Markov chain is not a uniform distribution. Let x be a random variable representing the number of cousins of a random protein in Ω_j (with the restriction that each mismatched segment is no longer than 2). The distribution of x is narrowly concentrated and has a small relative standard deviation.

A PTM in general modifies several amino acids, not all the 20 amino acids. In this case, a length n proteoform F is a modified proteoform of a length n protein P with PTMs $V = \{v_1, v_2, \ldots, v_k\}$ if (1) there are $n-k$ matched mass pairs in P and F and (2) the multiset of the mass differences of the remaining k mass pairs is the same as V, and (3) for each unmatched mass pair corresponding to an amino acid and a PTM (a mass shift), the PTM can modify the amino acid. In addition, we modify the definition of neighbor proteoforms in the greedy algorithm: two proteoforms F_1 and F_2 in $D(P, V)$ are neighbors if we can obtain F_2 from F_1 by shifting the position i of one PTM in F_1 to a new position j such that the amino acid at position j can be modified by the PTM. For the protein sequence P = GRMPKESK modified by a methylation and a phosphorylation, the proteoforms F_1 = GR[meth]MPKES[ph]K and F_2 = GRMPK[meth]ES[ph]K are neighbors, because F_2 can be obtained by shifting the position of the methylation site from the second amino acid R to the fifth amino acid K.

We define D_j as the set of protein sequences in D with a residue mass j, and d_j the size of D_j. Each sequence in D_j is randomly selected from the set Ω_j. Let $X(j, t, V)$ be a random variable representing the number of protein sequences P in D_j with $\text{PScore}(S, P, V) \geq t$. The p-values and E-values of PrSMs with various PTMs are estimated using the same method described previously.

Many proteoforms identified by top-down MS contain unexpected alterations. The proposed method can be extended to compute E-values and p-values of PrSMs containing unexpected alterations. For a PrSM with variable PTMs and an unexpected alteration with a mass shift x in $[-500, 500]$ Da, the proposed method is modified as follows: the mass shift x is considered as a variable PTM. An amino acid with a residue mass a can be modified by the PTM if $x + a > 0$.

5.3 Results

The proposed MCMC method was implemented in C++. All experiments were performed on a computer with an Intel Xeon E5-2637 3.50GHz CPU and 128 GB memory.

5.3.1 Evaluation of the greedy algorithm

The greedy algorithm in Figure 5.1 may fail to report correct similarity scores of protein spectrum matches with PTMs because its search space is limited. Large errors in estimated similarity scores will affect the accuracy of p-values reported by TopMCMC. We used the histone H4 data set to evaluate the accuracy of the greedy algorithm.

The human histone H4 protein sequence was downloaded from the UniProt database (version September 12, 2016) [72]. Acetylation, methylation, dimethylation, trimethylation, and phosphorylation were considered as variable PTMs. In a candidate proteoform, at most 10 variable PTMs were allowed and no unexpected mass shifts were allowed. Of the 3 256 spectra, the precursor masses of 1 112 matched (within 15 ppm) the molecular mass of a candidate proteoform of the histone H4 protein.

We computed two similarity scores: the P-score and G-score, for the match between each of the 1 112 spectra and the histone H4 protein with variable PTMs. For a protein

spectrum match (S, P, V) between a spectrum S and a protein P with a multiset V of PTMs, the G-score is an estimation of PScore(S, P, V) reported by the greedy algorithm. The PScore(S, P, V) is accurately computed by the graph alignment algorithm in TopMG. In the greedy algorithm, the error tolerance for fragment masses was 15 ppm. For each protein, the algorithm was performed 3 times with different initial random proteoforms (Step 1 in Figure 5.1), and the best score was reported.

The greedy method has a smaller search space and a shorter running time than the graph alignment method. The average running times of the greedy method and the graph alignment method on the 1 112 protein spectrum matches were 6 and 1 237 seconds, respectively. Because of the small search space of the greedy method, the G-score of a match was no larger than the P-score. We divided the 1 112 matches into four groups based on the number of PTMs in the best scoring proteoform reported by TopMG: $0-2$ PTMs, $3-5$ PTMs, $6-8$ PTMs, and $9-10$ PTMs. Figure 5.6 shows the scatter plots of the two scores of the matches in the four groups. The difference between the G-score and P-score of a PrSM increases as the number of PTMs increases. When the number of PTMs is no larger than 5, the difference between the two scores is 7.1 on average, and the standard deviation of the differences is 6.06. When the number of PTMs is larger than 5, the average and standard deviation of the differences between the two scores are increased to 10.7 and 10.32, respectively. In the MCMC method, a large variance in the differences significantly affects the accuracy of estimated p-values. The greedy method introduces more errors for matches with > 5 PTMs compared with those with ≤ 5 PTMs.

5.3.2 Evaluation based on p-values

The bipartite database strategy [96] was used to evaluate the accuracy of p-values reported by TopMCMC. In this strategy, query MS/MS spectra are searched against a bipartite protein database containing sample sequences and entrapment sequences. While the sample sequences are expected to be observed in the sample, the entrapment sequences are not. The p-values of matches between spectra and entrapment sequences should follow a uniform distribution. This property is used to assess the accuracy of methods that assign p-values to PrSMs.

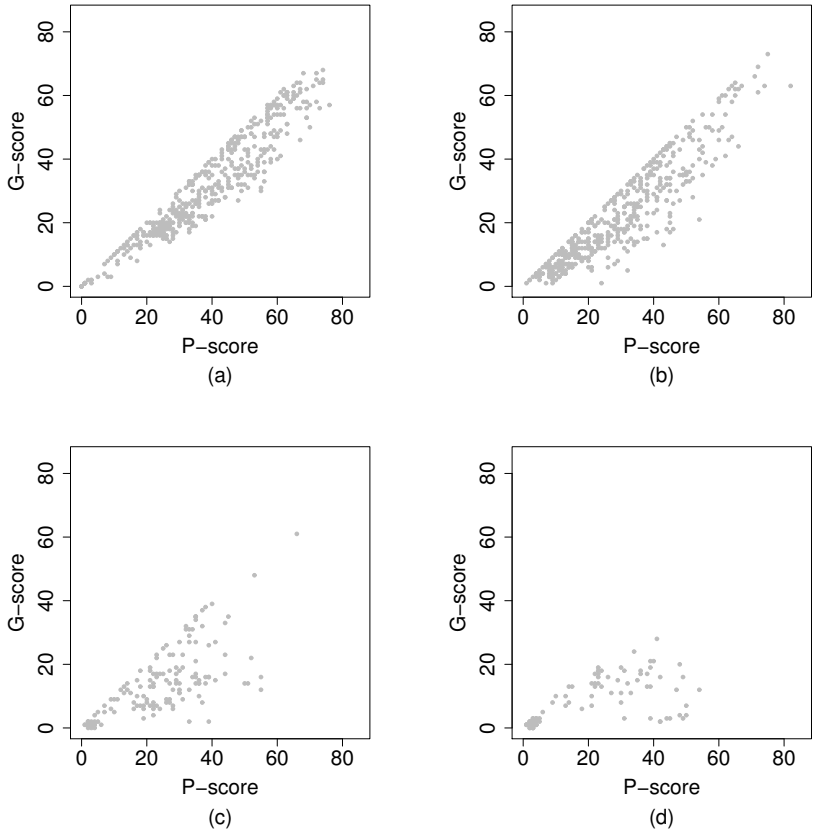

Figure 5.6: Scatter plots of the P-scores and G-scores of the 1 112 protein spectrum matches in the histone H4 data set with various numbers of PTMs: (a) $0-2$ PTMs; (b) $3-5$ PTMs; (c) $6-8$ PTMs; (d) $9-10$ PTMs.

We used the histone H3 data set to assess the accuracy of p-values estimated by TopMCMC. A bipartite database was constructed as follows. The 5 histone H3 protein sequences in the UniProt human proteome database (version September 12, 2016) were treated as sample sequences, and the sequences in the UniProt *Pyrococcus furiosus* proteome database (version February 4, 2017, 499 entries) entrapment ones. A previous study [97] demonstrated that *P. furiosus* proteins are a good choice for entrapment sequences because they have a long evolutionary distance with human sequences,

TopMG [1] was employed to search the spectra in the histone H3 data set against the bipartite database. Acetylation, methylation, dimethylation, trimethylation, and phosphorylation were considered as variable PTMs. The error tolerance for precursor and fragment masses was set to 15 ppm, at most 5 variable PTMs were allowed in an identified proteo-

form, and no unexpected mass shifts were allowed. Candidate PrSMs of a query spectrum can be divided into many types based on the number of PTMs and terminal truncations. TopMG reported a top scoring PrSM for each query spectrum and each PrSM type. The TopMCMC method was used to estimate p-values and E-values for the top scoring PrSMs and report one with the best E-value for each query spectrum. The greedy algorithm was used to speed up the estimation of P-Scores in TopMCMC. Of the 6 824 spectra, 2 638 were matched to proteoforms of the entrapment sequences.

By definition, the p-values of the entrapment PrSMs should follow a uniform distribution. One-sample Kolmogorov-Smirnov test was used to compute a D value (Kolmogorov-Smirnov statistic), a distance between the empirical distribution of the p-values reported by TopMCMC for the entrapment PrSMs and the uniform distribution over $[0, 1]$. The D value was 0.1874 with a p-value 2.2×10^{-16} (Figure 5.7), demonstrating that the empirical distribution and the uniform distribution are similar. Granholm et al. studied D values of scores reported by several commonly used tools for bottom-up mass spectral identification, such as SEQUEST (D value ≤ 0.03) and MS-GFDB (D value 0.21) [96]. The D values of SEQUEST and MS-GFDB are given for references, not for the comparison between TopMCMC and these tools. The average running time of TopMCMC for a PrSM was 2.13 seconds. The settings of the parameters $c_{max} = 10\,000$ and $T = 3$ were chosen to balance the running time and the accuracy of reported p-values.

We also compared cumulative relative frequencies of the p-values of the 2 638 PrSMs reported by TopMCMC and cumulative probabilities of the uniform distribution over $[0, 1]$ (Figure 5.8). If the cumulative relative frequency of the reported p-values for a value $x \in [0, 1]$ is larger than the cumulative probability of the uniform distribution for x, then the reported p-values in $[0, x]$ are underestimated. Figure 5.8 shows that TopMCMC underestimated the p-values in $[0, 0.7]$. The main reason is that rare events (PrSMs with high scores) might not be effectively sampled when the number of simulations (10 000 in the experiments) is not large enough.

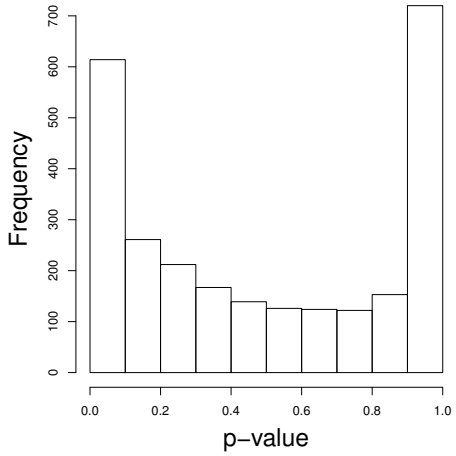

Figure 5.7: The histogram of p-values reported by TopMCMC for the 2 638 entrapment PrSMs reported from the histone H3 data set. The D value (Kolmogorov-Smirnov statistic) between the empirical distribution of the p-values and the uniform distribution over $[0, 1]$ is 0.1874.

5.3.3 Evaluation based on FDRs

We also evaluated the accuracy of TopMCMC using an false discovery rate (FDR)-based method [74, 98]. Given a list of query mass spectra, a target protein database, and an E-value cutoff t, the spectrum level FDR of identifications is estimated by two methods: one is by the target-decoy approach (TDA) [99], and the other by the eTDA estimator [98]. In the first method, the query mass spectra are searched against a concatenated target-decoy database for spectral identification, and the numbers of target and decoy identifications with an E-value better than t are used to estimate the FDR of the identifications. In the second method, each query mass spectrum is searched against the target database to find the best target PrSM, whose E-value is denoted by t_D. Then we compute the probability that the spectrum and a random decoy database, whose size is the same as the target database, have a PrSM with an E-value $< \min\{t, t_D\}$. That is, the decoy PrSM has an E-value better than t and than that of the best target PrSM. Such probabilities for all query spectra are summed up to obtain the expected number of decoy identifications, which is used to compute the expected FDR of identifications. The FDR estimated by the target-decoy approach is used as the gold standard. Because the computation of FDRs in the eTDA method is based on

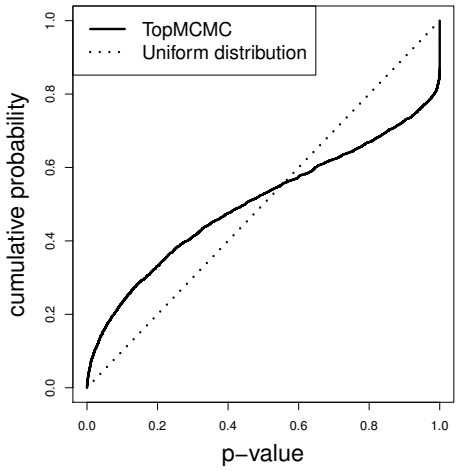

Figure 5.8: Comparison of the cumulative relative frequencies of the p-values reported by TopMCMC of the 2 638 entrapment PrSMs and the cumulative probabilities of the uniform distribution over $[0, 1]$. For each value x in $[0, 1]$, the cumulative relative frequency of the reported p-values in $[0, x]$ and the cumulative probability of the uniform distribution for x are plotted.

E-values reported by TopMCMC, a high similarity between the FDRs reported by the two methods demonstrates a high accuracy of the E-values reported by TopMCMC.

The EC data set was used in the evaluation. The UniProt EC proteome database (version September 12, 2016, 4 306 entries) was concatenated with a shuffled database of the same size. A two-step database search was performed to analyze the EC data set. First, the EC data set was searched against the EC proteome database using TopPIC [36] to quickly identify spectra generated from proteoforms without modifications or with one modification (some may contain terminal truncations). It is not necessary to use TopMCMC to estimate E-values of these identifications because they do not contain multiple modifications. One unexpected mass shift was allowed in an identified proteoform. With a 1% spectrum level FDR, a total of 1 920 PrSMs from 178 proteins were identified, including 470 PrSMs with unexpected mass shifts. Many mass shifts in the 470 PrSMs can be explained by common PTMs (Table 5.1). For example, mass shifts around 14 Da, which can be explained by methylation sites, were reported in 7 proteoforms from 4 proteins.

In the second step, these 1 450 spectra identified in the previous step were excluded, and the remaining 2 604 spectra (including the 470 spectra identified with mass shifts in the

Table 5.1: Common PTMs observed in the TopPIC identifications of EC data set.

PTM	Monoisotopic mass shift (Da)	Amino acids that can be modified	# proteins	# proteoforms
Acetylation	42.01056	R, K	8	9
Methylation	14.01565	R, K	4	7
Phosphorylation	79.96633	S, T, Y	1	1
Oxidation	15.99492	D, K, N, P, Y, R, C	9	9

previous step) were searched against the target-decoy EC database using TopMG. Because the mass shifts of acetylation, methylation, phosphorylation, and oxidation were observed in the first step of the analysis, they were treated as variable PTMs in TopMG. At most 5 variable PTM sites were allowed in a proteoform and no unexpected mass shifts were allowed.

A total of 303 and 86 PrSMs with an E-value smaller than 1 were reported from the target and decoy sequences, respectively. Since the FDR estimated by the target-decoy approach would be 0 when the cut-off E-value was below 1.11×10^{-4}, we only compared the FDRs for cut-off E-values greater than 1.11×10^{-4} (Figure 5.9). When the E-value cutoff is smaller than 0.1 ($-\log_{10}$(cutoff E-value) > 1), the FDRs estimated by the two methods are similar, and the FDRs estimated by the eTDA method are smaller than those by the TDA method, showing that E-values reported by TopMCMC are underestimated.

5.3.4 Discriminative capacity

We compared the TopMCMC method and the generating function approach [73, 74] on distinguishing correct identifications from incorrect ones using the MCF-7 data set. A human proteome database (version February 5, 2018, 20 303 entries) was downloaded from the UniProt database [72] and concatenated with a shuffled decoy database of the same size. Similar to the EC data set, a two-step database search was performed to analyze the MCF-7 data set. In the first step, all MCF-7 mass spectra were searched against the human target-decoy database using TopPIC, and the parameter settings were the same as the EC data analysis. With a 1% spectrum level FDR, TopPIC identified 615 PrSMs from

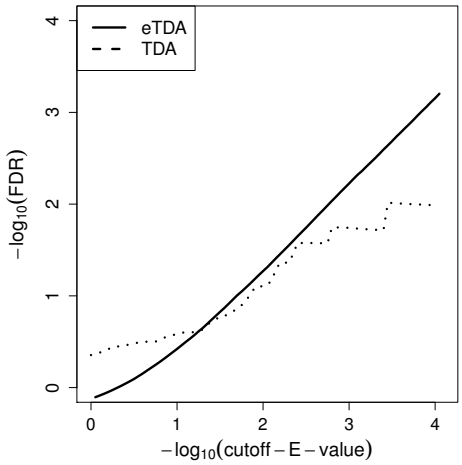

Figure 5.9: Comparison of the FDRs estimated by the TDA and eTDA methods for the PrSMs identified by TopMG from in the EC data set.

115 proteins, including 400 PrSMs without unexpected mass shifts. In the second step, the 400 spectra were excluded and the remaining 1 123 spectra were searched against the human target-decoy database using TopMG. The PTMs in Table 5.2 were considered as variable PTMs, and other parameter settings were the same as the EC data analysis. The TopMCMC method and the generating function method were incorporated into TopMG for E-value computation separately. TopMG coupled with TopMCMC is referred to as TopMG+MCMC, and TopMG coupled with the generating function method TopMG+GF.

Table 5.2: Common PTMs observed in the TopPIC identifications of MCF-7 data set.

PTM	Monoisotopic mass shift (Da)	Amino acids that can be modified	# proteins	# proteoforms
Acetylation	42.01056	R, K	5	5
Dimethylation	28.03130	R, K	2	2
Phosphorylation	79.96633	S, T, Y	8	15
Oxidation	15.99492	D, K, N, P, Y, R, C	3	3

With a 5% spectrum level FDR, TopMG+MCMC and TopMG+GF identified 161 and 133 PrSMs, respectively (Figure 5.10). TopMG+MCMC identified 21.1% more PrSMs than TopMG+GF, demonstrating that TopMCMC is better than the generating function

method in distinguishing correct identifications from incorrect ones. TopMG+GF missed many PrSMs because the implementation of the generating function method cannot accurately estimate E-values for PrSMs with multiple variable PTMs. TopMG+MCMC also missed 21 PrSMs identified by TopMG+GF. A possible reason is that the greedy method in TopMCMC introduced errors in the estimation of E-values of PrSMs with many variable PTMs. Most of the PrSMs (16 out of 21) missed by TopMG+MCMC have at least 4 variable PTMs. The running times of the two methods for E-value computations were similar: 380 seconds for TopMCMC and 375 seconds for the generating function method.

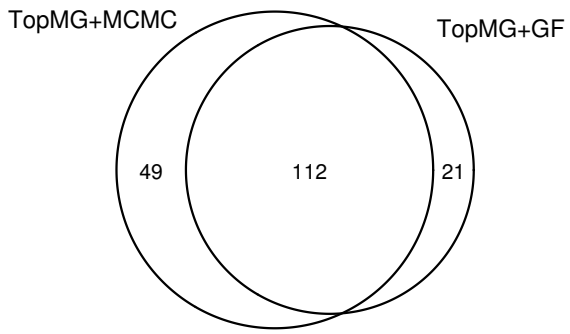

Figure 5.10: Comparison of the numbers of PrSMs identified by TopMG+MCMC and TopMG+GF from 1 123 spectra in the MCF-7 data set with a 5% spectrum level FDR.

5.4 Discussion

There are two main differences between TopMCMC and MS-DPR [93] although they use the same MCMC framework and oversampling method. First, while a sister of a peptide is obtained by changing two neighboring masses in MS-DPR, a sister of a protein is obtained by changing two masses or two substrings, which may be not neighbors, in TopMCMC. The definition of sister peptides in MS-DPR leads to smooth change of similarity scores after state transition, and that in TopMCMC leads to abrupt change of similarity scores. PrSMs identified by top-down MS often have a high similarity score. We need at least 30 transitions to move from a state with a score 0 to a state with a score 30. When the number of simulations is not large, the MCMC method may fail to find such a long path, resulting in inaccurate p-value estimation. Abrupt change of scores in TopMCMC can significantly

reduce the length of such a path, increasing the chance that states with high similarity scores are visited.

Second, the score of a peptide for a query spectrum in MS-DPR is the shared mass counting score between the spectrum and the peptide; the score of a protein in TopMCMC is the shared mass counting between the query spectrum and the best candidate proteoforms of the protein. Because the number of all candidate proteoforms grows exponentially with the number of PTM sites, a greedy method is used in TopMCMC to speed up the estimation of the similarity score.

TopMCMC is more accurate than the generating function method because it estimates protein-level probabilities, not proteoform-level probabilities. The generating function approach was designed to estimate E-values of matches between spectra and unmodified protein sequences. When it is extended to analyze PrSMs with variable PTMs, it can only report proteoform-level probabilities: the probability that a query spectrum and a random proteoform has a score no less than a threshold. Because many proteoforms of a protein are similar, the similarity scores of the query spectrum and these proteoforms are not independent. As a result, the generating function approach may have large errors in reported E-values. TopMCMC is capable of accurately estimating the protein-level probabilities: the probability that a query spectrum and the best scoring proteoform of a random protein has a score no less than a threshold, avoiding the errors caused by similar proteoforms. If users are interested in modification identification or proteoform characterization, modification identification scores or localization scores, such as the MIScore [100], can be reported as confidence scores of identified modifications.

The accuracy of p-values reported by TopMCMC is related to its number of simulations. While increasing the number of simulations will improve the accuracy, it also increases the running time. Experimental results demonstrated that TopMCMC achieved a good balance between the running time and the accuracy by setting c_{max} (the number of simulations) to 10 000 simulations and setting T (the number of rounds in oversampling factor estimation) to 3. The accuracy of reported p-values can be further improved by increasing the settings of c_{max} and T when a long running time is acceptable.

CHAPTER 6

CONCLUSION

The relatively small number of genes revealed by the Human Genome Project suggests a significant source of the complexity in human bodies is protein variation, which may come from genomic changes, *in vivo* proteolysis or PTMs. The accumulation of all these variations defines specific proteoforms [2], which govern various biological functions. In this dissertation, we present several novel algorithms for computational problems in complex proteoform identification using top-down MS/MS. All the proposed algorithms have been incorporated into TopMG, a complete software pipeline for complex proteoform identification. Experiments on simulated and real top-down MS/MS data sets showed TopMG can significantly increase the number of identifications compared with the existing methods. TopMG will facilitate the identification and quantification of clinically relevant proteoforms as well as the discovery of new protein biomarkers. Here, we summarize the main contributions of this dissertation and briefly discuss some future directions in the field of proteoform identification.

6.1 Summary

The contributions of this thesis involve several aspects of proteoform identification and are summarized as follows:

Approximate spectrum-based filtering algorithms Existing protein sequence filtering methods may fail when the target proteoform has more than two mass shifts or there are not enough consecutive fragment ions in the query spectrum. In the proposed ASF-RESTRICT and ASF-DIAGONAL algorithms, we address the problem by using a new strategy of incorporating the variable PTM information into MS/MS spectra, not the database sequences. The PTM information is used to remove variable PTMs in the match between the target sequence and the query spectrum.

Mass graph-based alignment algorithms We design a new data structure, called mass graphs, to effectively represent all possible proteoforms generated from the reference sequence with multiple variable PTMs and/or terminal truncations in one graph. One fundamental difference between the proposed structure and existing graph models in computational proteomics is to store amino acid residue masses as weights of edges, not of nodes. Using edge weights can significantly simplify the graphs in representing proteoforms with variable PTMs. Mass graph-based alignment algorithms are proposed for complex proteoform identification. Experiments on the simulated data set showed, even with 10 variable PTMs and terminal truncations, the proposed algorithms can still report more than 60% correct identifications.

Statistical significance estimation using MCMC In this dissertation, we design a new Markov chain model to represent proteins for top-down spectral interpretation, and a greedy algorithm is used to quickly estimate the similarity score between the query spectrum and a proteoform with multiple variable PTMs. The greedy algorithm and DPR sampling method together provide a fast method for estimating statistical significance of identified complex proteforms.

6.2 Future directions

This dissertation provides our solutions to several computational problems in proteoform identification. However, many fundamental issues in this field remain unresolved. In this section, we briefly discuss several promising research directions in this area.

Peak intensity A top-down MS/MS spectrum contains a list of peaks, each of which is represented as $(m/z, intensity)$. Almost every framework for MS-based proteoform identification ignores the intensity values. An exception is SQID, which incorporates intensity information in computing the scores of candidate peptides using bottom-up MS [101]. To our best knowledge, there is no software framework for top-down MS integrating the intensity information. Currently, TopMG uses the shared mass counting score, which only considers m/z values from each spectrum. The intensity value of a peak measures the abundance

of the corresponding fragment ion. However, due to the complexity of the fragmentation process, it is difficult to predict the intensity pattern and use it to assist proteoform identification. It is of great interest to investigate how to use the intensity information to improve the scoring function between mass spectra and proteoforms.

Retention time In a typical MS experiment, the protein samples are separated according to their hydrophobicity on an LC column. Then samples are ionized via ESI and subjected to mass spectrometry analysis. Besides the peak list, each spectrum also contains the retention time from the separation step. Recently, the prediction of retention time using protein sequences achieved very high accuracy [102, 103]. So we can predict retention time using database sequences and use the predicted retention time for sequence filtration. This filtering method can potentially improve the filtration efficiency and speed up the analysis workflow.

Spectral deconvolution Spectral deconvolution is usually one of the first steps in MS data analysis, and we assume the input of TopMG is mass spectra deconvoluted by TopFD. The deconvolution step determines the number of peaks and the noise level in each spectrum, which further affect the processing time and final results. A better deconvolution process can significantly reduce running time and improve proteoform identification results. Besides converting isotopomer envelopes into monoisotopic peaks, LC-MS feature detection is another critical step in spectral deconvolution. An LC-MS feature represents a group of isotopomer envelopes corresponding to the same proteoform across all charge states and retention times. Efficient LC-MS feature detection can significantly improve the proteoform identification and characterization results because the different spectra generated from the same proteoform can provide complementary information.

Proteogenomics Proteogenomics is an area of research combining proteomics and genomics [104]. One of the attractive features of proteogenomics research is to integrate information from both genomics and proteomics. Genomic and transcriptomic information is used to customize the protein database, and mass spectrometry is used to identify the novel proteoforms. On the other hand, the MS data provides the protein-level evidence for

genomic analysis. The complementary information from the same sample can provide a much deeper understanding of many problems in biomedical research.

BIBLIOGRAPHY

[1] Qiang Kou, Si Wu, Nikola Tolić, Ljiljana Paša-Tolić, Yunlong Liu, and Xiaowen Liu. A mass graph-based approach for the identification of modified proteoforms using top-down tandem mass spectra. *Bioinformatics*, 33:1309–1316, 2016.

[2] Lloyd M Smith, Neil L Kelleher, and Consortium for Top Down Proteomics. Proteoform: a single term describing protein complexity. *Nature Methods*, 10:186–187, 2013.

[3] Benjamin A Garcia, James J Pesavento, Craig A Mizzen, and Neil L Kelleher. Pervasive combinatorial modification of histone H3 in human cells. *Nature Methods*, 4:487–489, 2007.

[4] Nicolas L Young, Peter A DiMaggio, Mariana D Plazas-Mayorca, Richard C Baliban, Christodoulos A Floudas, and Benjamin A Garcia. High throughput characterization of combinatorial histone codes. *Molecular & Cellular Proteomics*, 8:2266–2284, 2009.

[5] Shahaf Peleg, Farahnaz Sananbenesi, Athanasios Zovoilis, Susanne Burkhardt, Sanaz Bahari-Javan, Roberto Carlos Agis-Balboa, Perla Cota, Jessica Lee Wittnam, Andreas Gogol-Doering, Lennart Opitz, Gabriella Salinas-Riester, Markus Dettenhofer, Hui Kang, Laurent Farinelli, Wei Chen, and André Fischer. Altered histone acetylation is associated with age-dependent memory impairment in mice. *Science*, 328:753–756, 2010.

[6] Xintong Dong, C Amelia Sumandea, Yi-Chen Chen, Mary L Garcia-Cazarin, Jiang Zhang, C William Balke, Marius P Sumandea, and Ying Ge. Augmented phosphorylation of cardiac troponin I in hypertensive heart failure. *Journal of Biological Chemistry*, 287:848–857, 2012.

[7] Emmanuel Boutet, Damien Lieberherr, Michael Tognolli, Michel Schneider, Parit Bansal, Alan J Bridge, Sylvain Poux, Lydie Bougueleret, and Ioannis Xenarios. UniProtKB/Swiss-Prot, the manually annotated section of the UniProt knowledge-

base: How to use the entry view. *Plant Bioinformatics: Methods and Protocols*, pages 23–54, 2016.

[8] John B Fenn, Matthias Mann, Chin Kai Meng, Shek Fu Wong, and Craig M Whitehouse. Electrospray ionization for mass spectrometry of large biomolecules. *Science*, 246(4926):64–71, 1989.

[9] Fred W McLafferty. Tandem mass spectrometry. *Science*, 214(4518):280–287, 1981.

[10] Scott A McLuckey. Principles of collisional activation in analytical mass spectrometry. *Journal of the American Society for Mass Spectrometry*, 3(6):599–614, 1992.

[11] J Mitchell Wells and Scott A McLuckey. Collision-induced dissociation (CID) of peptides and proteins. *Methods in Enzymology*, 402:148–185, 2005.

[12] Roman A Zubarev, Neil L Kelleher, and Fred W McLafferty. Electron capture dissociation of multiply charged protein cations. a nonergodic process. *Journal of the American Chemical Society*, 120(13):3265–3266, 1998.

[13] Ruedi Aebersold and Matthias Mann. Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207, 2003.

[14] Ruedi Aebersold and Matthias Mann. Mass-spectrometric exploration of proteome structure and function. *Nature*, 537(7620):347, 2016.

[15] Nagarjuna Nagaraj, Nils Alexander Kulak, Juergen Cox, Nadin Neuhauser, Korbinian Mayr, Ole Hoerning, Ole Vorm, and Matthias Mann. System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra hplc runs on a bench top orbitrap. *Molecular & Cellular Proteomics*, 11(3):M111–013722, 2012.

[16] Alicia L Richards, Alexander S Hebert, Arne Ulbrich, Derek J Bailey, Emma E Coughlin, Michael S Westphall, and Joshua J Coon. One-hour proteome analysis in yeast. *Nature Protocols*, 10(5):701, 2015.

[17] A. D. Catherman, O. S. Skinner, and N. L. Kelleher. Top down proteomics: facts and perspectives. *Biochemical and Biophysical Research Communications*, 445:683–93, 2014.

[18] Jungkap Park, Paul D Piehowski, Christopher Wilkins, Mowei Zhou, Joshua Mendoza, Grant M Fujimoto, Bryson C Gibbons, Jared B Shaw, Yufeng Shen, Anil K Shukla, Ronald J Moore, Tao Liu, Vladislav A Petyuk, Nikola Tolić, Ljiljana Paša-Tolić, Richard D Smith, Samuel H Payne, and Sangtae Kim. Informed-Proteomics: open-source software package for top-down proteomics. *Nature Methods*, 14:909–914, 2017.

[19] David M Horn, Roman A Zubarev, and Fred W McLafferty. Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *Journal of the American Society for Mass Spectrometry*, 11:320–332, 2000.

[20] Vlad Zabrouskov, Michael W Senko, Yi Du, Richard D Leduc, and Neil L Kelleher. New and automated MS^n approaches for top-down identification of modified proteins. *Journal of the American Society for Mass Spectrometry*, 16:2027–2038, 2005.

[21] Anoop M Mayampurath, Navdeep Jaitly, Samuel O Purvine, Matthew E Monroe, Kenneth J Auberry, Joshua N Adkins, and Richard D Smith. DeconMSn: a software tool for accurate parent ion monoisotopic mass determination for tandem mass spectra. *Bioinformatics*, 24:1021–1023, 2008.

[22] Paulo C Carvalho, Tao Xu, Xuemei Han, Daniel Cociorva, Valmir C Barbosa, and John R Yates III. YADA: a tool for taking the most out of high-resolution spectra. *Bioinformatics*, 25:2734–2736, 2009.

[23] Xiaowen Liu, Yuval Inbar, Pieter C Dorrestein, Colin Wynne, Nathan Edwards, Puneet Souda, Julian P Whitelegge, Vineet Bafna, and Pavel A Pevzner. Deconvolution and database search of complex tandem mass spectra of intact proteins: a combinatorial approach. *Molecular & Cellular Proteomics*, 9:2772–2782, 2010.

[24] Gordon W Slysz, Erin S Baker, Anuj R Shah, Navdeep Jaitly, Gordon A Anderson, and Richard D Smith. The DeconTools framework: an application programming

interface enabling flexibility in accurate mass and time tag workflows for proteomics and metabolomics. In *Proceedings of the 58th American Society Conference on Mass Spectrometry and Allied Topics*, 2010.

[25] Qiang Kou, Si Wu, and Xiaowen Liu. A new scoring function for top-down spectral deconvolution. *BMC Genomics*, 15:1140, 2014.

[26] Alexey I. Nesvizhskii. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *Journal of Proteomics*, 73:2092–123, 2010.

[27] Leonid Zamdborg, Richard D LeDuc, Kevin J Glowacz, Yong-Bin Kim, Vinayak Viswanathan, Ian T Spaulding, Bryan P Early, Eric J Bluhm, Shannee Babai, and Neil L Kelleher. ProSight PTM 2.0: improved protein identification and characterization for top down mass spectrometry. *Nucleic Acids Research*, 35(Web Server issue):W701–W706, 2007.

[28] Ari M Frank, James J Pesavento, Craig A Mizzen, Neil L Kelleher, and Pavel A Pevzner. Interpreting top-down mass spectra using spectral alignment. *Analytical Chemistry*, 80:2499–2505, 2008.

[29] Yihsuan S Tsai, Alexander Scherl, Jason L Shaw, C. Logan MacKay, Scott A Shaffer, Patrick R R Langridge-Smith, and David R Goodlett. Precursor ion independent algorithm for top-down shotgun proteomics. *Journal of the American Society for Mass Spectrometry*, 20:2154–2166, 2009.

[30] N. Murat Karabacak, Long Li, Ashutosh Tiwari, Lawrence J Hayward, Pengyu Hong, Michael L Easterling, and Jeffrey 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*, 8:846–856, 2009.

[31] Weiwei Tong, Roger Théberge, Giuseppe Infusini, David H Perlman, Catherine E Costello, and Mark E McComb. BUPID-top-down: database search and assignment

of top-down MS/MS data. In *Proceedings of the 57th American Society Conference on Mass Spectrometry and Allied Topics, Philadelphia, PA*, volume 31, 2009.

[32] Xiaowen Liu, Yakov Sirotkin, Yufeng Shen, Gordon Anderson, Yihsuan S Tsai, Ying S Ting, David R Goodlett, Richard D Smith, Vineet Bafna, and Pavel A Pevzner. Protein identification using top-down spectra. *Molecular & Cellular Proteomics*, 11:M111.008524, 2012.

[33] M. Bern, Y. J. Kil, and C. Becker. Byonic: advanced peptide and protein identification software. *Current Protocols in Bioinformatics*, Chapter 13:Unit 13.20, 2012.

[34] Li Li and Zhixin Tian. Interpreting raw biological mass spectra using isotopic mass-to-charge ratio and envelope fingerprinting. *Rapid Communications in Mass Spectrometry*, 27:1267–1277, 2013.

[35] Xiaowen Liu, Shawna Hengel, Si Wu, Nikola Tolić, Ljiljana Paša-Tolić, and Pavel A Pevzner. Identification of ultramodified proteins using top-down tandem mass spectra. *Journal of Proteome Research*, 12:5830–5838, 2013.

[36] Qiang Kou, Likun Xun, and Xiaowen Liu. TopPIC: a software tool for top-down mass spectrometry-based proteoform identification and characterization. *Bioinformatics*, 32:3495–3497, 2016.

[37] R. X. Sun, L. Luo, L. Wu, R. M. Wang, W. F. Zeng, H. Chi, C. Liu, and S. M. He. pTop 1.0: A high-accuracy and high-efficiency search engine for intact protein identification. *Analytical Chemistry*, 88:3082–90, 2016.

[38] Wenxuan Cai, Huseyin Guner, Zachery R Gregorich, Albert J Chen, Serife Ayaz-Guner, Ying Peng, Santosh G Valeja, Xiaowen Liu, and Ying Ge. MASH Suite Pro: A comprehensive software tool for top-down proteomics. *Molecular & Cellular Proteomics*, 15:703–714, 2016.

[39] Michael R Shortreed, Brian L Frey, Mark Scalf, Rachel A Knoener, Anthony J Cesnik, and Lloyd M Smith. Elucidating proteoform families from proteoform intact-mass and lysine-count measurements. *Journal of Proteome Research*, 15:1213–1221, 2016.

[40] Kaijie Xiao, Fan Yu, Houqin Fang, Bingbing Xue, Yan Liu, and Zhixin Tian. Accurate and efficient resolution of overlapping isotopic envelopes in protein tandem mass spectra. *Scientific Reports*, 5, 2015.

[41] Ryan T Fellers, Joseph B Greer, Bryan P Early, Xiang Yu, Richard D LeDuc, Neil L Kelleher, and Paul M Thomas. ProSight Lite: graphical software to analyze top-down mass spectrometry data. *Proteomics*, 15:1235–1238, 2015.

[42] Huseyin Guner, Patrick L Close, Wenxuan Cai, Han Zhang, Ying Peng, Zachery R Gregorich, and Ying Ge. MASH Suite: a user-friendly and versatile software interface for high-resolution mass spectrometry data interpretation and visualization. *Journal of the American Society for Mass Spectrometry*, 25:464–470, 2014.

[43] Zia Khan, Michael J Ford, Darren A Cusanovich, Amy Mitrano, Jonathan K Pritchard, and Yoav Gilad. Primate transcript and protein expression levels evolve under compensatory selection pressures. *Science*, 342(6162):1100–1104, 2013.

[44] Kathleen S Molnar, N Murat Karabacak, Joshua L Johnson, Qi Wang, Ashutosh Tiwari, Lawrence J Hayward, Stephen J Coales, Yoshitomo Hamuro, and Jeffrey N Agar. A common property of amyotrophic lateral sclerosis-associated variants: destabilization of the Cu/Zn superoxide dismutase electrostatic loop. *Journal of Biological Chemistry*, 2009.

[45] Matthew T Mazur, Helene L Cardasis, Daniel S Spellman, Andy Liaw, Nathan A Yates, and Ronald C Hendrickson. Quantitative analysis of intact apolipoproteins in human HDL by top-down differential mass spectrometry. *Proceedings of the National Academy of Sciences*, 107(17):7728–7733, 2010.

[46] Stacey R Oppenheimer, Deming Mi, Melinda E Sanders, and Richard M Caprioli. Molecular analysis of tumor margins by maldi mass spectrometry in renal carcinoma. *Journal of Proteome Research*, 9(5):2182–2190, 2010.

[47] Julia Chamot-Rooke, Guillain Mikaty, Christian Malosse, Magali Soyer, Audrey Dumont, Joseph Gault, Anne-Flore Imhaus, Patricia Martin, Mikael Trellet, Guilhem

Clary, et al. Posttranslational modification of pili upon cell contact triggers N. meningitidis dissemination. *Science*, 331(6018):778–782, 2011.

[48] Chad R Borges, Paul E Oran, Sai Buddi, Jason W Jarvis, Matthew R Schaab, Douglas S Rehder, Stephen P Rogers, Thomas Taylor, and Randall W Nelson. Building multidimensional biomarker views of type 2 diabetes based on protein microheterogeneity. *Clinical Chemistry*, 2011.

[49] Xintong Dong, C Amelia Sumandea, Yi-Chen Chen, Mary L Garcia-Cazarin, Jiang Zhang, C William Balke, Marius P Sumandea, and Ying Ge. Augmented phosphorylation of cardiac troponin i in hypertensive heart failure. *Journal of Biological Chemistry*, 287(2):848–857, 2012.

[50] Zhixin Tian, Nikola Tolić, Rui Zhao, Ronald J. Moore, Shawna M. Hengel, Errol W. Robinson, David L. Stenoien, Si Wu, Richard D. Smith, and Ljiljana Paša-Tolić. Enhanced top-down characterization of histone post-translational modifications. *Genome Biology*, 13:R86, 2012.

[51] Ioanna Ntai, Richard D. LeDuc, Ryan T. Fellers, Petra Erdmann-Gilmore, Sherri R. Davies, Jeanne Rumsey, Bryan P. Early, Paul M. Thomas, Shunqiang Li, Philip D. Compton, Matthew J C. Ellis, Kelly V. Ruggles, David Fenyö, Emily S. Boja, Henry Rodriguez, R Reid Townsend, and Neil L. Kelleher. Integrated bottom-up and top-down proteomics of patient-derived breast tumor xenografts. *Molecular & Cellular Proteomics*, 15:45–56, 2016.

[52] Philipp Mertins, Feng Yang, Tao Liu, D. R. Mani, Vladislav A. Petyuk, Michael A. Gillette, Karl R. Clauser, Jana W. Qiao, Marina A. Gritsenko, Ronald J. Moore, Douglas A. Levine, Reid Townsend, Petra Erdmann-Gilmore, Jacqueline E. Snider, Sherri R. Davies, Kelly V. Ruggles, David Fenyo, R. Thomas Kitchens, Shunqiang Li, Narciso Olvera, Fanny Dao, Henry Rodriguez, Daniel W. Chan, Daniel Liebler, Forest White, Karin D. Rodland, Gordon B. Mills, Richard D. Smith, Amanda G. Paulovich, Matthew Ellis, and Steven 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*, 13:1690–1704, 2014.

[53] Li Ding, Matthew J. Ellis, Shunqiang Li, David E. Larson, Ken Chen, John W. Wallis, Christopher C. Harris, Michael D. McLellan, Robert S. Fulton, Lucinda L. Fulton, Rachel M. Abbott, Jeremy Hoog, David J. Dooling, Daniel C. Koboldt, Heather Schmidt, Joelle Kalicki, Qunyuan Zhang, Lei Chen, Ling Lin, Michael C. Wendl, Joshua F. McMichael, Vincent J. Magrini, Lisa Cook, Sean D. McGrath, Tammi L. Vickery, Elizabeth Appelbaum, Katherine DeSchryver, Sherri Davies, Therese Guintoli, Li Lin, Robert Crowder, Yu Tao, Jacqueline E. Snider, Scott M. Smith, Adam F. Dukes, Gabriel E. Sanderson, Craig S. Pohl, Kim D. Delehaunty, Catrina C. Fronick, Kimberley A. Pape, Jerry S. Reed, Jody S. Robinson, Jennifer S. Hodges, William Schierding, Nathan D. Dees, Dong Shen, Devin P. Locke, Madeline E. Wiechert, James M. Eldred, Josh B. Peck, Benjamin J. Oberkfell, Justin T. Lolofie, Feiyu Du, Amy E. Hawkins, Michelle D. O'Laughlin, Kelly E. Bernard, Mark Cunningham, Glendoria Elliott, Mark D. Mason, Dominic M. Thompson Jr, Jennifer L. Ivanovich, Paul J. Goodfellow, Charles M. Perou, George M. Weinstock, Rebecca Aft, Mark Watson, Timothy J. Ley, Richard K. Wilson, and Elaine R. Mardis. Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature*, 464:999–1005, 2010.

[54] Shunqiang Li, Dong Shen, Jieya Shao, Robert Crowder, Wenbin Liu, Aleix Prat, Xiaping He, Shuying Liu, Jeremy Hoog, Charles Lu, Li Ding, Obi L. Griffith, Christopher Miller, Dave Larson, Robert S. Fulton, Michelle Harrison, Tom Mooney, Joshua F. McMichael, Jingqin Luo, Yu Tao, Rodrigo Goncalves, Christopher Schlosberg, Jeffrey F. Hiken, Laila Saied, Cesar Sanchez, Therese Giuntoli, Caroline Bumb, Crystal Cooper, Robert T. Kitchens, Austin Lin, Chanpheng Phommaly, Sherri R. Davies, Jin Zhang, Megha Shyam Kavuri, Donna McEachern, Yi Yu Dong, Cynthia Ma, Timothy Pluard, Michael Naughton, Ron Bose, Rama Suresh, Reida McDowell, Loren Michel, Rebecca Aft, William Gillanders, Katherine DeSchryver, Richard K. Wilson, Shaomeng Wang, Gordon B. Mills, Ana Gonzalez-Angulo, John R. Edwards,

Christopher Maher, Charles M. Perou, Elaine R. Mardis, and Matthew J. Ellis. Endocrine-therapy-resistant ESR1 variants revealed by genomic characterization of breast-cancer-derived xenografts. *Cell Reports*, 4:1116–1130, 2013.

[55] Roza Wojcik, Oluwatosin O Dada, Martin Sadilek, and Norman J Dovichi. Simplified capillary electrophoresis nanospray sheath-flow interface for high efficiency and sensitive peptide analysis. *Rapid Communications in Mass Spectrometry*, 24(17):2554–2560, 2010.

[56] Liangliang Sun, Guijie Zhu, Yimeng Zhao, Xiaojing Yan, Si Mou, and Norman J Dovichi. Ultrasensitive and fast bottom-up analysis of femtogram amounts of complex proteome digests. *Angewandte Chemie International Edition*, 52(51):13661–13664, 2013.

[57] Rachele A Lubeckyj, Elijah N McCool, Xiaojing Shen, Qiang Kou, Xiaowen Liu, and Liangliang Sun. Single-shot top-down proteomics with capillary zone electrophoresis-electrospray ionization-tandem mass spectrometry for identification of nearly 600 escherichia coli proteoforms. *Analytical Chemistry*, 89(22):12059–12067, 2017.

[58] D. Kessner, M. Chambers, R. Burke, D. Agus, and P. Mallick. Proteowizard: open source software for rapid proteomics tools development. *Bioinformatics*, 24:2534–6, 2008.

[59] Joel M Chick, Deepak Kolippakkam, David P Nusinow, Bo Zhai, Ramin Rad, Edward L Huttlin, and Steven P Gygi. A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nature Biotechnology*, 33:743–749, 2015.

[60] Matthias Mann and Matthias Wilm. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Analytical Chemistry*, 66:4390–4399, 1994.

[61] S. Tanner, H. Shu, A. Frank, L. C. Wang, E. Zandi, M. Mumby, P. A. Pevzner, and V. Bafna. InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Analytical Chemistry*, 77:4626–39, 2005.

[62] Ari Frank, Stephen Tanner, Vineet Bafna, and Pavel Pevzner. Peptide sequence tags for fast database search in mass-spectrometry. *Journal of Proteome Research*, 4:1287–1295, 2005.

[63] Xia Cao and Alexey I. Nesvizhskii. Improved sequence tag generation method for peptide identification in tandem mass spectrometry. *Journal of Proteome Research*, 7:4422–4434, 2008.

[64] David L. Tabb, Ze-Qiang Ma, Daniel B. Martin, Amy-Joan L. Ham, and Matthew C. Chambers. DirecTag: accurate sequence tags from peptide MS/MS through statistical scoring. *Journal of Proteome Research*, 7(9):3838–3846, Sep 2008.

[65] Sangtae Kim, Nitin Gupta, Nuno Bandeira, and Pavel A Pevzner. Spectral dictionaries integrating de novo peptide sequencing with database search of tandem mass spectra. *Molecular & Cellular Proteomics*, 8:53–69, 2009.

[66] Kyowon Jeong, Sangtae Kim, Nuno Bandeira, and Pavel A Pevzner. Gapped spectral dictionaries and their applications for database searches of tandem mass spectra. *Molecular & Cellular Proteomics*, 10:M110–002220, 2011.

[67] Fei Deng, Lusheng Wang, and Xiaowen Liu. An efficient algorithm for the blocked pattern matching problem. *Bioinformatics*, 31:532–538, 2014.

[68] Yufeng Shen, Nikola Tolić, Kim K Hixson, Samuel O Purvine, Gordon A Anderson, and Richard D Smith. De novo sequencing of unique sequence tags for discovery of post-translational modifications of proteins. *Analytical Chemistry*, 80:7742–7754, 2008.

[69] Xiaowen Liu, Alessandro Mammana, and Vineet Bafna. Speeding up tandem mass spectral identification using indexes. *Bioinformatics*, 28:1692–1697, 2012.

[70] Hao Chi, Kun He, Bing Yang, Zhen Chen, Rui-Xiang Sun, Sheng-Bo Fan, Kun Zhang, Chao Liu, Zuo-Fei Yuan, Quan-Hui Wang, Si-Qi Liu, Meng-Qiu Dong, and Si-Min He. pFind-Alioth: A novel unrestricted database search algorithm to improve the

interpretation of high-resolution MS/MS data. *Journal of Proteomics*, 125:89–97, 2015.

[71] Andy T Kong, Felipe V Leprevost, Dmitry M Avtonomov, Dattatreya Mellacheruvu, and Alexey I Nesvizhskii. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nature Methods*, 14:513–520, 2017.

[72] The UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Research*, 43(D1):D204–D212, 2015.

[73] Sangtae Kim, Nitin Gupta, and Pavel A. Pevzner. Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *Journal of Proteome Research*, 7(8):3354–3363, 2008.

[74] Xiaowen Liu, Matthew W Segar, Shuai Cheng Li, and Sangtae Kim. Spectral probabilities of top-down tandem mass spectra. *BMC Genomics*, 15(1):S9, 2014.

[75] S. Heber, M. Alekseyev, S. H. Sze, H. Tang, and P. A. Pevzner. Splicing graphs and EST assembly problem. *Bioinformatics*, 18 Suppl 1:S181–8, 2002.

[76] Y. Xing, A. Resch, and C. Lee. The multiassembly problem: reconstructing multiple transcript isoforms from est fragment mixtures. *Genome Research*, 14:426–41, 2004.

[77] S. Woo, S. W. Cha, G. Merrihew, Y. He, N. Castellana, C. Guest, M. MacCoss, and V. Bafna. Proteogenomic database construction driven from large scale RNA-seq data. *Journal of Proteome Research*, 13:21–8, 2014.

[78] S. Woo, S. W. Cha, S. Na, C. Guest, T. Liu, R. D. Smith, K. D. Rodland, S. Payne, and V. Bafna. Proteogenomic strategies for identification of aberrant cancer peptides using large-scale next-generation sequencing data. *Proteomics*, 14:2719–30, 2014.

[79] A. Frank and P. Pevzner. Pepnovo: De novo peptide sequencing via probabilistic network modeling. *Analytical Chemistry*, 77:964–973, 2005.

[80] L. He, X. Han, and B. Ma. De novo sequencing with limited number of post-translational modifications per peptide. *Journal of Bioinformatics and Computational Biology*, 11(4):1350007, 2013.

[81] S. Bhatia, Y. J. Kil, B. Ueberheide, B. T. Chait, L. Tayo, L. Cruz, B. Lu, 3rd Yates, J. R., and M. Bern. Constrained de novo sequencing of conotoxins. *Journal of Proteome Research*, 11(8):4191–200, 2012.

[82] N. Bandeira, D. Tsur, A. Frank, and P. A. Pevzner. Protein identification by spectral networks analysis. *Proceedings of the National Academy of Sciences USA*, 104:6140–5, 2007.

[83] M. S. Gelfand, A. A. Mironov, and P. A. Pevzner. Gene recognition via spliced sequence alignment. *Proceedings of the National Academy of Sciences USA*, 93:9061–6, 1996.

[84] Richard D LeDuc, Gregory K Taylor, Yong-Bin Kim, Thomas E Januszyk, Lee H Bynum, Joseph V Sola, John S Garavelli, and Neil L Kelleher. ProSight PTM: an integrated environment for protein identification and characterization by top-down mass spectrometry. *Nucleic Acids Research*, 32:W340–W345, 2004.

[85] William Stafford Noble and Michael J MacCoss. Computational and statistical analysis of protein mass spectrometry data. *PLoS Computational Biology*, 8:e1002296, 2012.

[86] Andrew Keller, Alexey I Nesvizhskii, Eugene Kolker, and Ruedi Aebersold. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analytical Chemistry*, 74(20):5383–5392, 2002.

[87] Rovshan G Sadygov and John R Yates. A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. *Analytical Chemistry*, 75(15):3792–3798, 2003.

[88] Rovshan G Sadygov, Hongbin Liu, and John R Yates. Statistical models for protein validation using tandem mass spectral data and protein amino acid sequence databases. *Analytical Chemistry*, 76(6):1664–1671, 2004.

[89] Alexey I Nesvizhskii and Ruedi Aebersold. Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem ms. *Drug Discovery Today*, 9(4):173–181, 2004.

[90] Sangtae Kim, Nikolai Mischerikow, Nuno Bandeira, J Daniel Navarro, Louis Wich, Shabaz Mohammed, Albert JR Heck, and Pavel A Pevzner. The generating function of CID, ETD, and CID/ETD pairs of tandem mass spectra: applications to database search. *Molecular & Cellular Proteomics*, 9(12):2840–2852, 2010.

[91] Reuven Y. Rubinstein and Dirk P. Kroese. *Simulation and the Monte Carlo Method*. Wiley-Interscience, 2 edition, 2007.

[92] Zsolt Haraszti and J Keith Townsend. The theory of direct probability redistribution and its application to rare event simulation. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 9:105–140, 1999.

[93] Hosein Mohimani, Sangtae Kim, and Pavel A Pevzner. A new approach to evaluating statistical significance of spectral identifications. *Journal of Proteome Research*, 12(4):1560–1568, 2013.

[94] Qiang Kou, Si Wu, and Xiaowen Liu. Systematic evaluation of protein sequence filtering algorithms for proteoform identification using top-down mass spectrometry. *Proteomics*, 18(3-4):1700306, 2018.

[95] Rajeev Motwani and Prabhakar Raghavan. *Randomized Algorithms*. Cambridge University Press, New York, NY, USA, 1995.

[96] Viktor Granholm, William Stafford Noble, and Lukas Käll. On using samples of known protein content to assess the statistical calibration of scores assigned to peptide-spectrum matches in shotgun proteomics. *Journal of Proteome Research*, 10(5):2671–2678, 2011.

[97] Marc Vaudel, Julia M Burkhart, Daniela Breiter, René P Zahedi, Albert Sickmann, and Lennart Martens. A complex standard for protein identification, designed by evolution. *Journal of Proteome Research*, 11(10):5065–5071, 2012.

[98] Nitin Gupta, Nuno Bandeira, Uri Keich, and Pavel A Pevzner. Target-decoy approach and false discovery rate: when things may go wrong. *Journal of the American Society for Mass Spectrometry*, 22(7):1111–1120, 2011.

[99] Joshua E. Elias and Steven P. Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods*, 4(3):207–214, Mar 2007.

[100] Qiang Kou, Binhai Zhu, Si Wu, Charles Ansong, Nikola Tolić, Ljiljana Paša-Tolić, and Xiaowen Liu. Characterization of proteoforms with unknown post-translational modifications using the miscore. *J Proteome Res*, 15(8):2422–2432, Aug 2016.

[101] Wenzhou Li, Li Ji, Jonathan Goya, Guanhong Tan, and Vicki H Wysocki. SQID: an intensity-incorporated protein identification algorithm for tandem mass spectrometry. *Journal of Proteome Research*, 10(4):1593–1602, 2011.

[102] Luminita Moruz and Lukas Käll. Peptide retention time prediction. *Mass Spectrometry Reviews*, 36(5):615–623, 2017.

[103] Chunwei Ma, Zhiyong Zhu, Jun Ye, Jiarui Yang, Jianguo Pei, Shaohang Xu, Chang Yu, Fan Mo, Bo Wen, and Siqi Liu. Retention time of peptides in liquid chromatography is well estimated upon deep transfer learning. *arXiv preprint arXiv:1711.00045*, 2017.

[104] Alexey I Nesvizhskii. Proteogenomics: concepts, applications and computational strategies. *Nature methods*, 11(11):1114, 2014.

CURRICULUM VITAE

Qiang Kou

Education

2013 - 2018 Doctor of Philosophy in Informatics,
Indiana University-Purdue University Indianapolis

2008 - 2012 Bachelor of Science in Biotechnology,
Sun Yat-sen University

Publications

1. Xiaojing Shen, **Qiang Kou**, Ruiqiong Guo, Zhichang Yang, Daoyang Chen, Xiaowen Liu, Heedeok Hong, Liangliang Sun. Native proteomics in discovery mode using size exclusion chromatography-capillary zone electrophoresis-tandem mass spectrometry. *Analytical chemistry*, 2018.

2. Elijah N McCool, Rachele A Lubeckyj, Xiaojing Shen, Daoyang Chen, **Qiang Kou**, Xiaowen Liu, Liangliang Sun. Deep top-down proteomics using capillary zone electrophoresis-tandem mass spectrometry: identification of 5700 proteoforms from the *Escherichia coli* proteome. *Analytical Chemistry*, 2018.

3. **Qiang Kou**, Si Wu, and Xiaowen Liu. Systematic evaluation of protein sequence filtering algorithms for proteoform identification using top-down mass spectrometry. *Proteomics*, 2018.

4. Runmin Yang, Daming Zhu, **Qiang Kou**, Poornima Bhat-Nakshatri, Harikrishna Nakshatri, Si Wu and Xiaowen Liu. A spectrum graph-based protein sequence filtering algorithm for proteoform identification by top-down mass spectrometry. *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2017.

5. Rachele A Lubeckyj, Elijah N McCool, Xiaojing Shen, **Qiang Kou**, Xiaowen Liu, Liangliang Sun. Single-shot top-down proteomics with capillary zone electrophoresis-electrospray ionization-tandem mass spectrometry for identification of nearly 600 *Escherichia coli* proteoforms. *Analytical Chemistry*, 2017.

6. **Qiang Kou**, Si Wu, Nikola Tolić, Ljiljana Paša-Tolić, Yunlong Liu and Xiaowen Liu. A mass graph-based approach for the identification of modified proteoforms using top-down

tandem mass spectra. *Bioinformatics*, 2017.

7. **Qiang Kou**, Likun Xun and Xiaowen Liu. TopPIC: a software tool for top-down mass spectrometry-based proteoform identification and characterization. *Bioinformatics*, 2016.

8. **Qiang Kou**, Binhai Zhu, Si Wu, Charles Ansong, Nikola Tolić, Ljiljana Paša-Tolić and Xiaowen Liu. Characterization of proteoforms with unknown post-translational modifications using the MIScore. *Journal of Proteome Research*, 2016.

9. Weijie Ding, **Qiang Kou**, Xueqin Wang, Qiuya Xu and Na You. Single-sample SNP detection by empirical Bayes method using next generation sequencing data. *Statistics and Its Interface*, 2015.

10. Murat Dundar, **Qiang Kou**, Baichuan Zhang, Yicheng He and Bartek Rajwa. Simplicity of kmeans versus deepness of deep learning: A case of unsupervised feature learning with limited data. *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, 2015.

11. **Qiang Kou**, Si Wu, and Xiaowen Liu. A new scoring function for top-down spectral deconvolution. *BMC Genomics*, 2014.

12. Xiaowen Liu, Lennard JM Dekker, Si Wu, Martijn M. Vanduijn, Theo M. Luider, Nikola Tolić, **Qiang Kou**, Mikhail Dvorkin, Sonya Alexandrova, Kira Vyatkina, Ljiljana Paša-Tolić and Pavel A. Pevzner. De novo protein sequencing by combining top-down and bottom-up tandem mass spectra. *Journal of Proteome Research*, 2014.

13. Na You, Peng Mou, Ting Qiu, **Qiang Kou**, Huaijin Zhu, Yuexi Chen, Xueqin Wang. Gene Expression Network Reconstruction by LEP Method Using Microarray Data. *The Scientific World Journal*, 2012