Published in final edited form as:
J Proteome Res. 2013 December 6; 12(12): 5830–5838. doi:10.1021/pr400849y.

Identification of ultramodified proteins using top-down tandem mass spectra

Xiaowen Liu[*,†,‡], **Shawna Hengel**[¶], **Si Wu**[¶], **Nikola Tolić**[¶], **Ljiljana Pasa-Tolić**[¶], and **Pavel A. Pevzner**[§]

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

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

[¶]Environmental Molecular Sciences Laboratory Pacific Northwest National Laboratory

[§]Department of Computer Science and Engineering, University of California, San Diego

Abstract

Post-translational modifications (PTMs) play an important role in various biological processes through changing protein structure and function. Some *ultramodified* proteins (like histones) have multiple PTMs forming *PTM patterns* that define the functionality of a protein. While bottom-up mass spectrometry (MS) has been successful in identifying *individual* PTMs within short peptides, it is unable to identify PTM patterns spreading along entire proteins in a coordinated fashion. In contrast, top-down MS analyzes intact proteins and reveals PTM patterns along the entire proteins. However, while recent advances in instrumentation have made top-down MS accessible to many laboratories, most computational tools for top-down MS focus on proteins with few PTMs and are unable to identify complex PTM patterns. We propose a new algorithm, MS-Align-E, that identifies both expected and unexpected PTMs in ultramodified proteins. We demonstrate that MS-Align-E identifies many proteoforms of histone H4 and benchmark it against the currently accepted software tools.

Introduction

Post-translational modifications (PTMs) affect protein structure and function. In some proteins, the function of the protein is determined by a *combination* of multiple PTM sites (*PTM pattern*) rather than individual PTMs at specific sites. We refer to proteins with many PTM sites as *ultramodified* proteins. For example, histones often have multiple PTM sites with various PTM types such as acetylation, methylation, and phosphorylation. Specifically for histones, the PTM patterns define their gene regulatory functions[1,2] through the "combinatorial histone code".[3,4] PTM patterns in histones are part of the epigenetic mechanisms that are now being linked to several human diseases. However, revealing PTM patterns in histones has proven to be a challenge. As Garcia and colleagues wrote in a recent review: "The ability to detect combinatorial histone PTMs is now much easier than it has been before, but the most difficult issue with these analyses still remains: deconvolution of the data".[5] Highly complex top-down spectra of histones feature multiple ion series that are either shared and unique to the multiple proteoforms. These spectra have to be decoded for

[*]Corresponding author Phone: +1-317-278-7613. Fax: +1-317-278-9201. xwliu@iupui.edu.
Present address Department of BioHealth Informatics, Indiana University-Purdue University Indianapolis, 719 Indiana Avenue, Indianapolis, IN 46202, USA.

Suuporting information available: This material is available free of charge via the Internet at http://pubs.acs.org.

revealing the histone PTM space and deriving rules governing the combinatorial histone code.

PTMs are often classified into *expected* and *unexpected* referring to the types of PTMs that are commonly and rarely observed (on specific proteins). For example, with respect to histones, acetylation, methylation, and phosphorylation represent expected PTMs, while carbamylation may represent an unexpected PTM. We emphasize that by expected PTMs we mean expected PTM *types* rather than PTM *sites*. Expected PTM types are often referred to as “variable PTMs” in peptide identification tools.

Bottom-up database search tools offer a variety of algorithms for searching for both expected[6] and unexpected[7,8] PTMs. However, while bottom-up mass spectrometry (MS) has been successful in identifying some PTM sites, it is not well suited for identification of complex PTM patterns. Because bottom-up MS is based on digesting proteins into short peptides, PTMs identified are restricted to *individual* peptides, lacking information on how many protein isoforms are present (i.e. how the combination of modified/unmodified peptide sequences are put back together). Even if all peptides within a protein and all PTMs within each peptide were identified, the ability to identify PTM patterns would still be lacking because the correlations between PTMs located on different peptides are lost (Fig. 1). Moreover, bottom-up MS rarely provides full coverage of proteins by identified peptides: a typical shotgun proteomics study (with a single protease like trypsin) provides on average about 25% coverage for proteins.[9] It implies that many PTMs may remain below the radar of bottom-up proteomics. Middle-down proteomics[10,11] identifies PTM sites on longer peptides and thus takes an intermediate position between bottom-up and top-down approaches with respect to identifying PTM patterns, however there is still a gap between intact proteoforms and digestion products.

Over the last several years, applications of top-down MS have significantly expanded due to the recent progress in MS instrumentation and protein separation. The widely available commercial mass spectrometers are now capable of analyzing short proteins with molecular weight up to 30 kDa.[12] However, software tools for analyzing ultramodified proteins by top-down MS have not kept pace with rapid developments in top-down MS technology.

The main challenge in analysis of ultramodified proteins lies in the complexity of these proteins. A ultramodified protein may have a large number of possible proteoforms.[13] For instance, based on the UniProt[14] flat file, histone H4 has more than 26 billion potential proteoforms. Researchers have made significant effort to separate individual proteoforms.[3,4,15,16] However, multiplexed tandem mass spectra still exist in top-down liquid chromatography-tandem mass spectrometry (LC/MS/MS) analysis of ultramodified proteins due to the similarity of proteoforms.[11,13] Data analysis of these top-down tandem mass spectra can be categorized into two problems: (1) Identification of the most abundant proteoform in a tandem mass spectrum, and (2) identification and qualification of multiple proteoforms in a multiplexed tandem mass spectrum. The second problem has been well covered in the studies of several groups. DiMaggio *et al.* and Baliban *et al.* employed integer-linear optimization to identify and qualify multiple proteoforms in multiplexed spectra.[10,11] Guan *et al.* used non-redundant ions to classify peptides or proteoforms into independent configurations, the associated dependent configurations, and unsupported configurations, and qualify independent configurations in multiplexed spectra.[13] In this paper, we focus on identification of the most abundant proteoform in a tandem mass spectrum.

Existing top-down protein identification tools score Protein-Spectrum-Matches (PrSMs) using various scoring functions $Score(P, S)$, where (P, S) refers to a PrSM formed by a protein *P* and a spectrum *S*. The simplest scoring function (called the "shared peak count") counts the number of peaks in the spectrum *S* "explained" by the protein *P*, i.e., the number of shared monoisotopic peaks between *S* and the *theoretical spectrum* of *P*. Given a PrSM (P^*, *S*) between a proteoform P^* of a protein *P* with PTMs and a spectrum *S*, the shared peak count is the number of shared monoisotopic peaks between *S* and the theoretical spectrum of P^*.

Given an unmodified protein *P*, a set of expected PTM types Ω, and an integer *F*, we define *ProteinDB*(*P*, Ω, *F*) as the set of all proteoforms of *P* with exactly *F* expected PTM sites. Since the size of *ProteinDB*(*P*, Ω, *F*) increases exponentially with an increase in *F*, exploring all proteoforms in this database becomes computationally intractable, particularly when the set of expected PTM types is large. This motivates the following *Expected PTM Identification* (EPI) problem: given a top-down spectrum *S*, an unmodified protein *P*, an integer *F*, and a set of expected PTM types Ω, find a proteoform P^* of the protein *P* with *F* expected PTM sites such that *Score*(P^*, *S*) is maximized among all proteoforms in *ProteinDB*(*P*, Ω, *F*).

MS-Align-E (Mass Spectral ALIGNment for Expected PTMs) solves the EPI problem and is further extended for identifying proteins with both expected and unexpected PTMs. Even in the case of closely located sites of expected PTMs, MS-Align-E is capable of identifying correct PTM patterns. We tested MS-Align-E on a top-down MS/MS data set from histone H4. We further compared the resulting PrSMs with those reported by MS-TopDown[17] and ProSightPC.[18]

Methods

MS-Align-E uses the spectral alignment[17] to find PrSMs and the generating function approach[19] to compute the *E*-values of these PrSMs. The key part of the generating function approach is the assumption that amino acids have integer masses.[19] However, rounding amino acid masses into integers introduces errors. These rounding errors reduce after rescaling by 0.9995.[20–22] While the scaling constant 0.9995 proved to be useful for bottom-up peptide identification, the rounding errors remain too large, even after rescaling, for highly accurate top-down spectra. MS-Align-E uses a scaling constant 274.335215 (e.g. mass(*G*) = $57.021464 \times 274.335215 = 15642.995586 \approx 15643$) that reduces the rounding error to 2.5 parts per million (ppm). We thus assume that masses of all amino acids are integers (the mass of an amino acid *r* is referred to as *mass*(*r*)).

A protein $B = r_1 r_2 \ldots r_m$ is a sequence of amino acids. The mass of a protein prefix $r_1 r_2 \ldots r_k$ is $b_k = \sum_{i=1}^{k} mass(r_i)$. We will find it convenient to represent a protein *B* as a sequence of its prefix masses $b_0 < b_1 < \ldots < b_m$ (we assume $b_0 = 0$). The molecular mass of protein *B* is $\sum_{i=1}^{m} mass(r_i) + mass(H_2O)$, where *mass*($H_2O$) is the (rounded) mass of a water molecule.

A tandem mass spectrum (MS/MS) generated from a protein is represented by a precursor mass and a list of peaks. The precursor mass corresponds to the molecular mass of the protein and each peak, represented as (*m*/*z*, *intensity*), corresponds to a fragment ion of the protein. The values *m*/*z* and *intensity* are the mass-to-charge ratio and the abundance of the fragment ion, respectively. In preprocessing of top-down MS/MS spectra, *m*/*z* values are usually converted into neutral masses of fragment ions by deconvolution algorithms.[23,24] Most of the neutral masses correspond to either protein prefixes or protein suffixes. The list of neutral masses can be further converted to a list of *prefix residue masses (PRMs)* corresponding to the masses of protein prefixes.[25] For a collision-induced dissociation (CID) spectrum with a precursor mass *M*, the PRM spectrum is generated as follows: (1) two masses 0 and $M - mass(H_2O)$ are added to the PRM spectrum (the mass $M - mass(H_2O)$ equals to the sum of the masses of all residues in the protein); (2) for each neutral mass x extracted from the experimental spectrum, two masses x and $M - x$ are added to the PRM spectrum. If mass x corresponds to a protein suffix (prefix), then mass $M - x$ corresponds to a protein prefix (suffix). Similar to discretization of amino acid masses, the precursor masses and the PRMs are discretized resulting in PRM spectra with integer mass values.

In contrast to bottom-up peptide identification tools that benefit from information about peak intensities, the existing top-down protein identification algorithms hardly use information about peak intensities (except for filtering out low intensity peaks). While in this paper we also ignore peak intensities, all proposed algorithms can be easily generalized to incorporate peak intensities. We represent a PRM spectrum A with a precursor mass M simply as a list of ordered integers $a_0 < a_1 < \dots < a_n$, where $a_0 = 0$ and $a_n = M - mass(H_2O)$.

The mass difference between an amino acid residue with a PTM and the unmodified same residue is the *mass shift* of the PTM. A PTM with a mass shift s on the ith residue in B transforms it into $b_0, b_1, \dots, b_i + s, \dots, b_m + s$. The mass shifts of all PTMs are discretized in the same way as PRMs. Let $S_1 = \{s_1, s_2, \dots, s_k\}$ be the set of mass shifts corresponding to the expected PTM types in the EPI problem. The (composite) *mass shift* of several expected PTM sites is the sum of their mass shifts. The set of mass shifts of all combinations of f expected PTM sites is defined recursively as $S_f = \{s | s = u + v, u \in S_1 \text{ and } v \in S_{f-1}\}$, for $f = 2, 3, \dots$. For example, if $S_1 = \{14, 42\}$, then $S_2 = \{28, 56, 84\}$ and $S_3 = \{42, 70, 98,126\}$. The *modification number* of an integer s is the minimum number f satisfying $s \in S_f$. For example, when $S_1 = \{14, 42\}$, the composite mass shift 84 is present in three sets S_2, S_4, and S_6 since $84 = 42 + 42 = 42 + 14 + 14 + 14 = 14 + 14 + 14 + 14 + 14 + 14$. The modification number of 84 is 2. We also define $mod(0) = 0$ and $mod(s) = \infty$ if s cannot be partitioned into a sum of integers from S_1.

Typically, a PTM type modifies only several types of amino acids rather than all 20 standard amino acids. For example, phosphorylation is observed on amino acids S, T, and Y, but not on A. To simplify the presentation, we first consider a rather unrealistic case when each expected PTM type can modify all 20 amino acids. We will later describe how MS-Align-E restricts each expected PTM type to some specific amino acids that it can modify.

Spectral alignment

Given sequences of integers $A = a_0, a_1, \dots, a_n$ and $B = b_0, b_1, \dots, b_m$, the *grid* of A and B is defined as a two dimensional grid within a rectangle formed by four points $(0, 0)$, $(b_m, 0)$, $(0, -a_n)$, $(b_m, -a_n)$.[17] The grid has $(n + 1)(m + 1)$ *matching points* $p_{i,j} = (b_j, -a_i)$. We refer to the upper leftmost matching point $(0, 0)$ and the lower rightmost matching point $(b_m, -a_n)$ as the *source* and the *sink*, respectively. Given matching points $p_{i,j}$ and $p_{i',j'}$, we say $p_{i',j'} < p_{i,j}$ if $i' < i$ and $j' < j$. We construct a *grid graph* with vertices corresponding to matching points and directed edges from matching points $p_{i',j'}$ to $p_{i,j}$ if $p_{i',j'} < p_{i,j}$. The grid graph has $O(n \cdot m)$ vertices and $O(n^2 \cdot m^2)$ edges.

The *mass shift* of an edge from vertex (matching point) $p_{i',j'}$ to vertex $p_{i,j}$ is defined as $(a_i - b_j) - (a_{i'} - b_{j'})$. An edge is called a *diagonal edge* if its mass shift is zero, and a *shift edge* otherwise. The diagonal edges are represents by (−45°) diagonal segments. An *alignment* between sequences A and B is a path from the source to the sink in the grid graph. We assign *scores* to the vertices in the grid graph and define the score of an alignment (path) as the total score of its vertices. Below we assume that every vertex in the grid graph has score 1. An *optimal alignment* is an alignment with the maximum score.

As an example, consider a protein B =GSTGRTK and its modified version B^* =GS[+160]T[-30]GRT[-30]K with 3 PTMs. The grid for these proteins (represented as sequences $B = \{0, 57, 144, 245, 302, 458, 559, 687\}$ and $B^* = \{0, 57, 304, 375, 432, 588, 659, 787\}$) is shown in Fig. 2(a). The alignment shown in Fig. 2(a) represents every unmodified (modified) amino acid as a diagonal (shift) edge. The score of the alignment is simply the number of vertices in the alignment path (length of the protein plus 1).

Fig. 2(b) shows the grid in the case when the protein B^* is substituted by its spectrum A. Compared to B^*, the spectrum A has two missing masses 304 and 432, and a noise mass 482. As a result, the optimal alignment in Fig. 2(b) differs from the alignment in Fig. 2(a): the missing mass 384 results in substituting two consecutive shift edges by a single one, while the missing mass 432 results in substituting two consecutive diagonal edges by a single one.

When A and B correspond to a spectrum and a peptide, we refer to the grid and alignment between them as their *spectral grid* and *spectral alignment*, correspondingly. Diagonal edges in a spectral alignment correspond to segments of B matched to spectrum A without PTMs; shift edges correspond to segments of B with PTMs. The *modification number* of an edge is defined as the modification number of its mass shift (e.g., diagonal edges have modification number 0). The modification number of an edge from $p_{i',j'}$ to $p_{i,j}$ is denoted by $mod(p_{i',j'} \to p_{i,j})$. A shift edge from $p_{i',j'}$ to $p_{i,j}$ is *valid* if its modification number $x \le F$ and $x \le j - j'$. The condition $x \le j - j'$ guarantees that for a shift edge with modification number x, there exist at least x modified residues in the protein supporting the mass shift. A spectral alignment is *valid* if all its shift edges are valid. The *modification number* of a spectral alignment is the sum of the modification numbers of its shift edges. A spectral alignment between A and B with modification number F is *optimal* if it has the maximum score among all alignments with modification number F. It is easy to check that a path shown in Fig. 2(b) is an optimal valid alignment with modification number 3. Since a valid spectral alignment with a modification number F corresponds to a proteoform with F PTM sites,[26] the EPI problem is reduced to the following graph-theoretical problem:

Expected PTM spectral alignment (EPSA) problem

Given a spectrum $A = \{a_0, a_1, \dots , a_n\}$, a protein $B = \{b_0, b_1, \dots , b_m\}$, an integer F, a set of mass shifts S_1 corresponding to expected PTMs, find an optimal valid spectral alignment of A and B with the modification number F.

To solve the EPSA problem one can use the *parametric dynamic programming* algorithm (similar to the generating function approach[19]) for finding a longest path in a spectral grid graph with a given number of modifications. However, the running time of the longest path algorithm is proportional to the number of edges in the spectral grid graph (proportional to $n^2 \cdot m^2$ making this algorithm prohibitively time consuming). Pevzner *et al.*, 2000, 2001[26,27] described an *equivalent transformation* of the spectral grid graph that greatly reduces the number of edges in the graph while preserving an optimal spectral alignment path. We develop a similar approach, EPSA algorithm, for top-down spectra (See the supplementary material for details). Let $\mathcal{S} = \{0\} \cup S_1 \cup \dots \cup S_F$ and $T = \min\{|\mathcal{S}|, (n + 1)(m + 1)\}$. The running time of the EPSA algorithm is proportional to $n \cdot m \cdot T$.

From spectral grids to diagonal grids

A mass spectrum A of protein B contains fragment ions corresponding to *some* but not necessarily all cleavage sites of B. As a result, the spectral alignment in Fig. 2(b) deteriorates as compared to Fig. 2(a). However, given the set of (composite) mass shifts $\mathcal{S}$, one can construct a set of prefix residue masses corresponding to *all* putative cleavage sites of protein B (and to "restore" the quality of spectral alignment) as follows.

A −45° line l passing the spectral grid at point (x, y) is called a *diagonal line* of $offset(l) = -x - y$. For example, a diagonal line starting at the left vertical border of the grid at $(0, -10)$ has offset 10. Similarly to the standard grid formed by crossing $(n + 1)$ horizontal lines with $(m + 1)$ vertical lines (originated from spectrum $A = \{a_0, \dots , a_n\}$ and protein $B = \{b_0, \dots , b_m\}$), we form a *diagonal grid* by crossing $|\mathcal{S}|$ diagonal lines with $(m+1)$ vertical lines. For each $s \in \mathcal{S}$, there exists a diagonal line with offset s contributing to the diagonal grid (Fig. 2(c)). The intersection of a diagonal line and a vertical line is called a *diagonal point* (there are $|\mathcal{S}| \cdot (m+1)$ diagonal points in the diagonal grid). Let $l_0, l_1, \dots , l_{|\mathcal{S}|-1}$ be the diagonal lines ordered in the increasing order of $offset(l_0) < offset(l_1) < \dots < offset(l_{|\mathcal{S}|-1})$. The diagonal point of a crossing line l_i and a vertical line corresponding to mass b_j is denoted by $q_{i,j}$.

The *diagonal grid graph* (or simply *diagonal graph*) is defined similarly to the grid graph. The vertex set of the diagonal graph consists of all diagonal points. Score 1 is assigned to vertices in the diagonal grid if they are present in the spectral grid (all other vertices are assigned score 0). The set of edges in the diagonal graph is redefined (as compared to the spectral grid graph) by only connecting vertices located on *consecutive* vertical lines in the diagonal grid. Specifically, a vertex (diagonal point) $q_{i,j}$ is connected with a vertex $q_{i+1,j'}$ by an edge if the difference between the offsets of diagonal lines l_j and $l_{j'}$ is either 0 (i.e., connecting consecutive vertices on the same diagonal line) or in set S_1.

A *diagonal alignment* is defined as an alignment (path) in the diagonal graph (Fig. 2(c)). Each valid path in the spectral grid graph has a corresponding path in the diagonal grid graph (all shift edges have a modification number 1). Edges with modification number larger than 1 in the spectral grid graph correspond to paths (formed by edges with modification number 1) in the diagonal graph. As Fig. 2(c) illustrates, the diagonal alignment improves as compared to the spectral alignment in Fig. 2(b) and now looks like the protein-protein alignment in Fig. 2(a). The EPSA problem in the spectral grid graph is reduced to the following problem in the diagonal graph:

Expected PTM diagonal alignment (EPDA) problem

Given a spectrum $A = \{a_0, a_1, \dots , a_n\}$, a protein $B = \{b_0, b_1, \dots , b_m\}$, an integer F, a set of mass shifts S_1 corresponding to expected PTMs, find an optimal diagonal alignment of A and B with F shift edges in the diagonal graph.

We designed an EPDA algorithm for the problem (See the supplementary material for details). The running time of the EPDA algorithm is proportional to $m \cdot F \cdot |\mathcal{S}|$, a significant speed-up compared to the EPSA algorithm.

Typically, a PTM type modifies only several types of amino acids rather than all 20 amino acids. Restricting PTMs to a subset of amino acids can be naturally modeled in the framework of the diagonal graph. Since every shift edge in the diagonal graph corresponds to a specific amino acid in the protein, we simply remove shift edges whose shift values are not present in the list of allowed PTMs for the amino acid.

Identifying spectra with both expected and unexpected PTMs

The spectral alignment algorithms can be modified to identify proteins with both expected and unexpected PTMs.[17] However, the complexity of the resulting algorithm is proportional to $n \cdot m \cdot T \cdot F_e \cdot F_u$, where $T = \min\{(n + 1)(m + 1), |\mathcal{S}|\}$, and F_e and F_u are the numbers of expected and unexpected PTM sites, respectively. Since this algorithm is too slow in practice, we propose a fast heuristic algorithm for identifying proteins with both expected and unexpected PTMs (See the supplementary material). To identify protein isoforms truncated at N- or C-terminus, a local alignment algorithm[28] is used. *E*-values of identified PrSMs are computed using a generating function approach.[19]

Mass spectrometry experiment

The proposed method was tested using a histone H4 MS/MS data set. Primary normal human dermal fibroblasts (NHDFs) were obtained from Lonza (Allendale, NJ) and grown in FGM-2 media (Lonza). Cells were harvested from 10 confluent 150 mm plates by trypsin digestion and washed 3 times in PBS. Core histones were purified using a histone purification kit according to the manufacturer's instructions (Active Motif; Carlsbad, CA), and precipitated overnight by the addition of perchloric acid to a final concentration of 4%. Following centrifugation, pellets were washed 2 times with 4% perchloric acid, 2 times with acetone containing 0.2% HCl, and 2 times with 100% acetone. Air dried pellets were resuspended in 200 μL H_2O and stored at -80° C until use.

Core histones (10 μg) were analyzed using a custom histone 2D RP-HILIC system coupled directly to a LTQ Orbitrap Velos (Thermo Scientific, Waltham, MA). Histone H4 was isolated in the first dimension of separation. Electrospray ionization voltage, 4.5 kV, was applied by connecting the end of the HILIC column to a 20 μm inner diameter chemically etched capillary emitter with a PEEK union (**See** Fig. 2 in the supplementary materialfor the LC curve); while a voltage was applied through a metal union downstream of the analyte. Histone H4 acquisitions were performed in the Orbitrap with nominal resolving power of 60,000. FTMS MS and MSn AGC target values were 10^6 and 5×10^5, respectively. Two micro scans were summed for all acquired spectra. Fragmentation of the top five most intense precursor ions, isolated with a 3 *m/z* window, was performed by alternating Collision-Induced Dissociation (CID) and Electron Transfer Dissociation (ETD) for the same precursor ion. Dynamic exclusion was implemented with exclusion duration of 200 s and an exclusion list size of 150. MS/MS was only performed on species with charge states greater than 4. In total, 1,626 CID and 1,626 ETD spectra were acquired.

Results

We implemented MS-Align-E in Java and tested it on the top-down MS/MS data set of histone H4. The experiments were carried out on a desktop PC with 3.4 GHz CPU (Intel Core i7-3770) and 16 GB memory.

Identification of proteoforms from ultramodified histone H4

All MS/MS spectra were deconvoluted using MS-Deconv;[24] precursor ions are deconvoluted within a window of 3 *m/z* in MS spectra. MS-Align-E was used to align the deconvoluted spectra with the histone H4 protein sequence. The error tolerances for precursor ions and fragment ions were set as 15 ppm. Five PTM types were treated as expected ones (Table 1); maximum 10 expected PTM sites and 1 unexpected PTM site were allowed. Because the mass of deamidation is about 1 Da and deisotoping of top-down tandem mass spectra often introduces ±1 Da error in precursor and fragment masses, it is common for protein identification tools to report erroneous identification of deamidation sites. Therefore, we excluded deamidation from the list of expected PTMs.

The running time of MS-Align-E was about 505 minutes (with computing *E*-values). With *E*-value cut off 0.01[1], MS-Align-E identified 629 spectra[2]. These results can provide hints

[1] The target/decoy approach was used to estimate false discovery rate of the identified PrSMs, but no PrSMs with an *E*-value ≤ 0.01 were reported from the shuffled decoy protein database.

to help identify and functionally characterize different proteoforms of histone H4. Many identified spectra have more than 3 expected PTM sites (Fig. 3). When one unexpected PTM site is allowed, several expected or unexpected PTM sites might be combined to an unexpected PTM site with a large mass shift. Thus, the proteoforms with one unexpected PTM sites tend to have less expected PTM sites compared with those without unexpected PTM sites.

Comparison with MS-TopDown

MS-TopDown[17] was downloaded from http://proteomics.ucsd.edu/Software.html. Only ETD spectra were used for comparison since MS-TopDown is hard-coded for ETD spectra. The ETD spectra were analyzed by MS-TopDown and MS-Align-E using the parameter setting in the previous section. The running time of MS-TopDown and MS-Align-E was 17 and 88 minutes, respectively (since MS-TopDown does not compute *E*-values, we ran MS-Align-E without computing *E*-values). While MS-TopDown is faster than MS-Align-E, it does not consider combinations of several expected PTM sites, thus limiting its ability to find proteoforms.

Since MS-TopDown does not report *E*-values, the number of matched fragment ions was used to rank identified PrSMs. MS-TopDown and MS-Align-E identified 327 and 456 PrSMs with at least 10 matched fragment ions, respectively (Fig. 4). In most cases, the proteoform reported by MS-Align-E had more matched fragment ions than that reported by MS-TopDown (for the same spectrum). For example, the proteoform reported by MS-Align-E for the spectrum of scan number 2,858 had 75 matched fragment ions while the one reported by MS-TopDown had 33 matched fragment ions. When the number of matched fragment ions is small, it is possible that no fragment ions support the cleavage sites between two expected PTM sites. As a result, the two PTM sites can be found only if combinations of multiple expected PTM sites are included in spectral alignment. MS-TopDown failed to identify many PrSMs because it did not consider combinations of several expected PTM sites.

Comparison with ProSightPC

While ProSightPC[18] is capable of identifying proteoforms in a protein mixture, MS-Align-E is proposed for identifying proteoforms of a purified protein. In this paper, we only compare the performance of the two tools when a single purified protein is analyzed. To identify proteoforms in a protein mixture, one needs to use ProSightPC or other software tools, such as MS-Align+.[28]

ProSightPC computes *E*-values of identified PrSMs based on the size of the target protein database and a Poisson distribution of three parameters: the number of fragment ions, the number of matched fragment ions, and the probability of an observed fragment ion matching a random theoretical fragment ion. MS-Align-E uses a generating function approach[19] to estimate *E*-values of identified PrSMs. Because ProSightPC and MS-Align-E report different *E*-values for the same PrSM, it is not fair to compare the number of PrSMs identified by the two tools using the same cutoff for *E*-values. Alternatively, the number of matched fragment ions was used to compare PrSMs identified by the two tools. All PrSMs with at least 10 matched fragment ions were reported and compared.

The annotated human proteoform database with 10,535,964 proteoforms was downloaded from ftp://prosightpc.northwestern.edu/2012_06/Eukaryotes/Homo sapients/. Because the

[2] The number of proteoforms identified by MS-Align-E was not reported because MS-Align-E did not use a proteoform database and some spectra did not have enough fragment masses to localize all PTMs.

data set contains a large number of tandem mass spectra, the high-throughput processing of ProSightPC was employed for data analysis, which is based on the “absolute mass” mode. ProSightPC also provides the Δ m mode which can be combined with the “absolute mass” mode to identify proteoforms with PTMs not included in the annotated proteoform database. By coupling with MS-Deconv,[24] two search modes of ProSightPC were tested: (1) the “absolute mass” mode and (2) the “absolute mass” mode combined with the Δm mode. All tests were performed on ProSightPC 2.0.[18]

In the “absolute mass” mode, the error tolerances for precursor ions and fragment ions were set as 2.2 Da and 15 ppm. The spectra were searched against all proteoforms in the annotated human proteoform database with PTM types in Table 1. The running time of ProSightPC was about 12 minutes. Using the same parameter setting as ProSightPC[3], MS-Align-E was applied to analyze the same data set. The running time of MS-Align-E was 577 minutes. ProSightPC is faster than MS-Align-E since its search space (all proteoforms in histone H4 in the annotated proteoform database) is far smaller than the search space of MS-Align-E (all combinations of expected PTMs in histone H4).

ProSightPC and MS-Align-E identified 1,029 spectra (from 100 proteoforms) and 1, 031 spectra with at least 10 matched fragment ions, respectively. Of the 1029 spectra identified by ProSightPC, 1024 spectra were reported by both tools. ProSightPC identified 5 PrSMs missed by MS-Align-E. Manual analysis of the 5 spectra showed that most of the spectra had a relatively large error in the precursor mass. Because MS-Align-E used precursor masses to compute prefix residue masses (PRMs) for spectral alignment, it failed to identify the spectra because the inaccurate precursor masses introduced large errors into PRMs. On the contrary, ProSightPC does not use PRM spectrum in proteoform identification and is capable of identifying correct proteoforms even if precursor masses are not accurate.

When a spectrum was identified by the two tools, two different proteoforms may be reported for the spectrum. Let P and M be the numbers of matched fragment ions reported by ProSightPC and MS-Align-E for a spectrum, respectively. Of the 1024 spectra, ProSightPC reported a better proteoform than MS-Align-E ($P > M$) for 135 spectra (Fig. 5). The difference $P - M \geq 5$ for 25 spectra. Manual analysis showed that MS-Align-E failed to identify the proteoforms due to the same reason as it missed 5 PrSMs identified ProSightPC. For 312 spectra, M is greater than P (Fig. 5(a)). This difference is large ($M - P \geq 5$) for 14 spectra. Fig. 6 illustrates the case when the “absolute mass” mode of ProSightPC reports an erroneous PrSM between the spectrum of scan number 2,062 and a known proteoform from ProSightPC database (The MS and MS/MS spectra are in the supplementary material). While this PrSM is high-scoring, the correct PrSM (found by MS-Align-E) explains many more fragmentation sites and has a much higher score. Two more examples (the spectra of scan numbers 3,199, and 3,507) that most likely represent the error in the “absolute mass” of ProSightPC are given in the supplementary material. The reason is that the “absolute mass” mode of ProSightPC (not combined with the Δm mode) tends to use known proteoforms to explain spectra originating from unknown proteoforms.

To test the “absolute mass” mode combined with the Δm mode, the error tolerance for precursor ions was set to 100 Da and other parameters were the same to the previous test. ProSightPC reported 1,135 PrSMs with at least 10 matched fragment ions. MS-Align-E missed 124 spectra identified by ProSightPC. Of the 1,011 spectra were identified by both tools, ProSightPC reported a PrSM with more matched fragment ions than MS-Align-E for

[3] For MS-Align-E, an error tolerance for precursor masses was used which is comparable to the default setting 2.2 Da in ProSightPC: A deconvoluted monoisotopic precursor mass m matches a theoretical precursor ion with neutral mass m' if the minimum error among the five mass pairs $(m - 2, m')$, $(m - 1, m')$, (m, m'), $(m + 1, m')$, $(m + 2, m')$ is not greater than 15 ppm.

740 spectra (Fig. 5(b)). ProSightPC outperformed MS-Align-E on proteoform identification because it combined the Δ*m* mode and a relatively complete annotated proteoform database of histone H4. In most cases, MS-Align-E missed some matched fragment ions due to the errors in precursor masses. The combined mode of ProSightPC avoided this problem by correcting the errors of precursor masses using molecular masses of known proteoforms. The combined mode also corrected most of erroneous PrSMs reported by the "absolute mass" mode. For example, it reported the correct proteoform with one unexpected PTM site (the location of the unexpected PTM site was not given) for the spectrum of scan number 2, 062 because the annotated proteoform database contains a proteoform of histone H4 with an N-terminal methionine excision and four PTMs: three acetylation sites on the second residue 'S', the 13th residue 'K', and the 17th residue 'K', and one dimethylation site on the 21th residue 'K'. The proteoform misses only one PTM (methylation on the 56th residue 'R') compared with the proteoform reported by MS-Align-E. However, the combined mode of ProSightPC missed 20 spectra identified by MS-Align-E (See the spectra of scan numbers 3,199 and 3,507 in the supplementary material). One possible reason is that the annotated proteoform database of ProSightPC does not contain a proteoform of histone H4 with an N-terminal methionine excision and one PTM: acetylation on the second residue 'S', which missed only one unexpected PTM compared to the proteoforms reported by MS-Align-E.

We remark that the performance of ProSightPC may be compromised when the annotated proteoform database is not complete. In contrast, MS-Align-E searched against only the unmodified form of histone H4 and achieved comparable performance to ProSightPC. When the annotated proteoform database is not available, MS-Align-E can identify many novel proteoforms.

Discussion

There are two main approaches to solving the EPI problem. The "virtual database" approach (proposed by Neil Kelleher's group and implemented in ProSightPC[18]) compares each spectrum against the "virtual database" with the goal to find the best scoring PrSM.[11,18] This approach faces a combinatorial explosion when the number of PTM sites is large. A ultramodified protein may have a very large number of potential proteoforms making it impractical to generate a "virtual database" containing all its proteoforms. The number of proteoforms explodes even further in searches for both expected and unexpected PTMs. Another limitation of the "virtual database" approach is its performance depends on the completeness of the annotated proteoform database. Similar to ProSightPC, PILOT_PTM[10] and the algorithm proposed by Guan *et al.*[13] enumerate all possible proteoforms for a given molecular mass in identification of proteoforms. Even when the molecular mass is fixed, the number of potential proteoforms of a ultramodified protein is still very large due to combinatorial explosion.

To avoid combinatorial explosion, the *spectral alignment* algorithms for top-down protein identification find the best-scoring PrSM without explicitly exploring all proteoforms in the virtual database in the case-by-case fashion.[17,28] However, the existing spectral alignment approaches, while working well for identification of proteins with a relatively small number of PTM sites (e.g., up to 3-4), were not designed for identification of ultramodified proteins like histones. First, they are primarily aimed at unexpected PTMs and the capabilities remain limited in the case of searches for both expected and unexpected PTMs. For example, due to limitations of the scoring functions, they tend to interpret two closely located expected PTM sites with masses *a* and *b* as a single unexpected PTM with mass *a*+*b*. Another limitation of the existing spectral alignment tools is that they require evidence for each PTM in the form of a "diagonal" in the spectral alignment matrix (See[28]). When there are no fragmentation sites between two consecutive PTM sites along the protein, such diagonals may not exist, preventing the spectral alignment algorithms from solving the EPI problem. This situation is quite common for histones since PTM sites in histones are often closely located to each other.

Acknowledging that the “virtual database” approach is useful for identification of known proteoforms and unknown proteoforms (using the Δ*m* mode of ProSightPC), we emphasize that its performance depends on an annotated proteoform database. The Δ*m* mode of ProSightPC can identify one unexpected PTM site not included in the database, but the localization of the PTM still needs manual analysis. When a protein lacks the database of its proteoforms, many experiments and analyses are required to create an annotated proteoform database to increase the number of identified proteoforms. This process may be time consuming and increase the cost of research. In addition, the “absolute mass” mode of ProSightPC may report erroneous identifications when the proteoforms are not in the annotated proteoform database.

MS-Align-E addresses this limitation of ProSightPC since it does not rely on a “virtual database”. It is an efficient tool for identifying proteoforms with multiple PTMs automatically, especially when a large number of spectra are analyzed. The main disadvantage of MS-Align-E is that it may report incorrect proteoforms due to errors in precursor or fragment masses, multiplexing spectra, and PTMs with similar mass shifts. Therefore, combining MS-Align-E, ProSightPC and manual annotation tools, such as the single protein mode in ProSightPC, will further improve the accuracy of identification of proteoforms with multiple PTMs.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by National Institutes of Health Grant P-41-RR024851 and a startup fund provided by Indiana University-Purdue University Indianapolis. Portions of this work were performed in the William R. Wiley Environmental Molecular Sciences Laboratory (EMSL), a DOE BER national scientific user facility located on the campus of PNNL in Richland, Washington. PNNL is a multi-program national laboratory operated by Battelle for the DOE under Contract DE-AC05-76RLO1830. We thank Dr. David Stenoien for providing the core histone mixtures.

References

1. Cosgrove MS, Wolberger C. How does the histone code work? Biochemistry and Cell Biology. 2005; 83:468–476. [PubMed: 16094450]
2. Strahl BD, Allis CD. The language of covalent histone modifications. Nature. 2000; 403:41–45. [PubMed: 10638745]
3. Garcia BA, Pesavento JJ, Mizzen CA, Kelleher NL. Pervasive combinatorial modification of histone H3 in human cells. Nature Methods. 2007; 4:487–9. [PubMed: 17529979]
4. Young NL, DiMaggio PA, Plazas-Mayorca MD, Baliban RC, Floudas CA, Garcia B. High throughput characterization of combinatorial histone codes. Molecular & Cellular Proteomics. 2009; 8:2266–84. [PubMed: 19654425]
5. Britton LMP, Gonzales-Cope M, Zee BM, Garcia BA. Breaking the histone code with quantitative mass spectrometry. Expert Review of Proteomics. 2011; 8:631–643. [PubMed: 21999833]
6. Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature. 2003; 422:198–207. [PubMed: 12634793]
7. Tsur D, Tanner S, Zandi E, Bafna V, Pevzner PA. Identification of post-translational modifications by blind search of mass spectra. Nature Biotechnology. 2005; 23:1562–1567.

8. Na S, Bandeira N, Paek E. Fast multi-blind modification search through tandem mass spectrometry. Molecular & Cellular Proteomics. 2012; 11:M111.010199. [PubMed: 22186716]

9. de Godoy LMF, Olsen JV, de Souza GA, Li G, Mortensen P, Mann M. Status of complete proteome analysis by mass spectrometry: SILAC labeled yeast as a model system. Genome Biology. 2006; 7:R50. [PubMed: 16784548]

10. Baliban RC, DiMaggio PA, Plazas-Mayorca MD, Young NL, Garcia BA, Floudas CA. A novel approach for untargeted post-translational modification identification using integer linear optimization and tandem mass spectrometry. Molecular & Cellular Proteomics. 2010; 9:764–779. [PubMed: 20103568]

11. DiMaggio PA J, Young NL, Baliban RC, Garcia BA, Floudas CA. A mixed integer linear optimization framework for the identification and quantification of targeted post-translational modifications of highly modified proteins using multiplexed electron transfer dissociation tandem mass spectrometry. Molecular & Cellular Proteomics. 2009; 8(11):2527–43. [PubMed: 19666874]

12. Tran JC, Zamdborg L, Ahlf DR, Lee JE, Catherman AD, Durbin KR, Tipton JD, Vellaichamy A, Kellie JF, Li M, Wu C, Sweet SMM, Early BP, Siuti N, Leduc RD, Compton PD, Thomas PM, Kelleher NL. Mapping intact protein isoforms in discovery mode using top-down proteomics. Nature. 2011; 480:254–258. [PubMed: 22037311]

13. Guan S, Burlingame AL. Data processing algorithms for analysis of high resolution msms spectra of peptides with complex patterns of posttranslational modifications. Mol Cell Proteomics. 2010; 9(5):804–10. [PubMed: 19955076]

14. Consortium U. Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Research. 2012; 40(Database issue):D71–5. [PubMed: 22102590]

15. Udeshi ND, Compton PD, Shabanowitz J, Hunt DF, Rose KL. Methods for analyzing peptides and proteins on a chromatographic timescale by electron-transfer dissociation mass spectrometry. Nature Protocols. 2008; 3(11):1709–17.

16. Tian Z, Zhao R, Tolic N, Moore RJ, Stenoien DL, Robinson EW, Smith RD, Pasa-Tolic L. Two-dimensional liquid chromatography system for online top-down mass spectrometry. Proteomics. 2010; 10(20):3610–20. [PubMed: 20879039]

17. Frank AM, Pesavento JJ, Mizzen CA, Kelleher NL, Pevzner PA. Interpreting top-down mass spectra using spectral alignment. Analytical Chemistry. 2008; 80:2499–2505. [PubMed: 18302345]

18. Zamdborg L, LeDuc RD, Glowacz KJ, Kim YB, Viswanathan V, Spaulding IT, Early BP, Bluhm EJ, Babai S, Kelleher NL. ProSight PTM 2.0: improved protein identification and characterization for top down mass spectrometry. Nucleic Acids Research. 2007; 35:W701–W706. [PubMed: 17586823]

19. Kim S, Gupta N, Pevzner PA. Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. Journal of Proteome Research. 2008; 7:3354–3363. [PubMed: 18597511]

20. Kim S, Gupta N, Bandeira N, Pevzner PA. Spectral dictionaries: Integrating *de novo* peptide sequencing with database search of tandem mass spectra. Molecular & Cellular Proteomics. 2009; 8:53–69. [PubMed: 18703573]

21. Taylor JA, Johnson RS. Sequence database searches via *de novo* peptide sequencing by tandem mass spectrometry. Rapid Communications in Mass Spectrometry. 1997; 11:1067–75. [PubMed: 9204580]

22. Bern M, Cai Y, Goldberg D. Lookup peaks: a hybrid of *de novo* sequencing and database search for protein identification by tandem mass spectrometry. Analytical Chemistry. 2007; 79:1393–400. [PubMed: 17243770]

23. Horn DM, Zubarev RA, McLafferty FW. Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. Journal of the American Society for Mass Spectrometry. 2000; 11:330–332.

24. Liu X, Inbar Y, Dorrestein PC, Wynne C, Edwards N, Souda P, Whitelegge JP, Bafna V, Pevzner PA. Deconvolution and database search of complex tandem mass spectra of intact proteins: A combinatorial approach. Molecular & Cellular Proteomics. 2010; 9:2772–2782. [PubMed: 20855543]

25. Tanner S, Shu H, Frank A, Wang LC, Zandi E, Mumby M, Pevzner PA, Bafna V. InsPecT: Identification of posttranslationally modified peptides from tandem mass spectra. Analytical Chemistry. 2005; 77:4626–4639. [PubMed: 16013882]

26. Pevzner PA, Dančík V, Tang CL. Mutation-tolerant protein identification by mass spectrometry. Journal of Computational Biology. 2000; 7:777–787. [PubMed: 11382361]

27. Pevzner PA, Mulyukov Z, Dancik V, Tang CL. Efficiency of database search for identification of mutated and modified proteins via mass spectrometry. Genome Research. 2001; 11:290–9. [PubMed: 11157792]

28. Liu X, Sirotkin Y, Shen Y, Anderson G, Tsai YS, Ting YS, Goodlett DR, Smith RD, Bafna V, Pevzner PA. Protein identification using top-down spectra. Molecular & Cellular Proteomics. 2012:M111.008524.

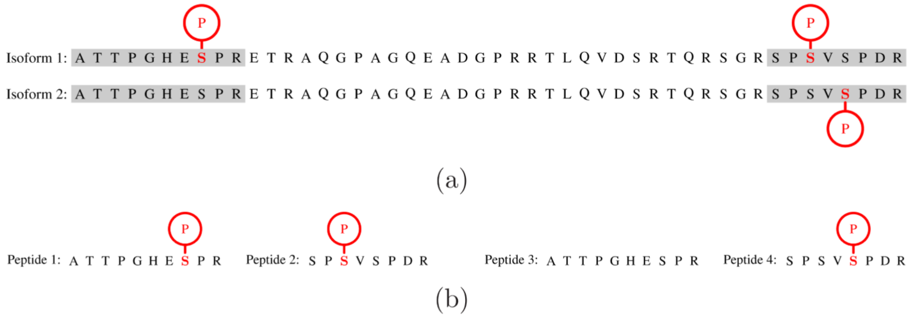

Isoform 1: A T T P G H E S P R E T R A Q G P A G Q E A D G P R R T L Q V D S R T Q R S G R S P S V S P D R
Isoform 2: A T T P G H E S P R E T R A Q G P A G Q E A D G P R R T L Q V D S R T Q R S G R S P S V S P D R
P
(a)
Peptide 1: A T T P G H E S P R
Peptide 2: S P S V S P D R
Peptide 3: A T T P G H E S P R
Peptide 4: S P S V S P D R
(b)

Figure 1. Bottom-up MS lacks the ability to identify complex PTM patterns

(a) Two proteoforms of a protein with phosphorylation sites coexist in the sample (P represents phosphorylation). (b) Bottom-up MS identifies four peptides (shaded regions in the proteoforms) resulting in up to 4 putative proteoforms. However, it is unable to answer the question which of these putative proteoforms are present in the sample.

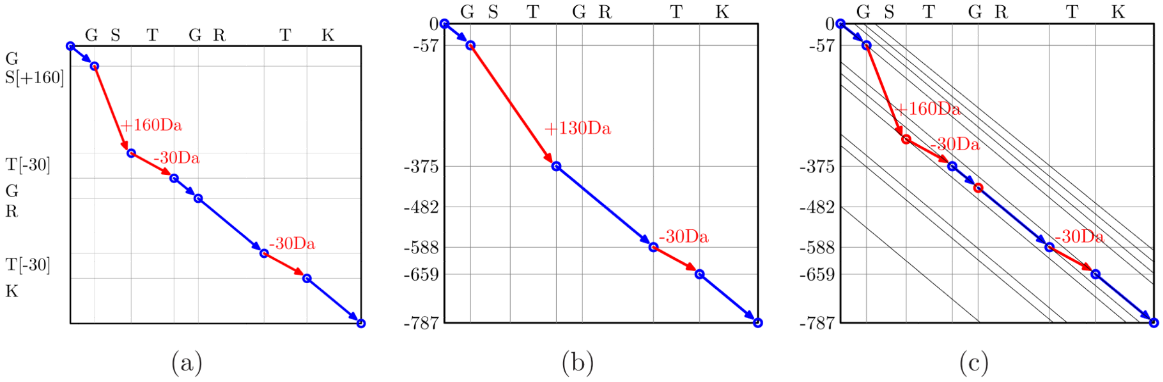
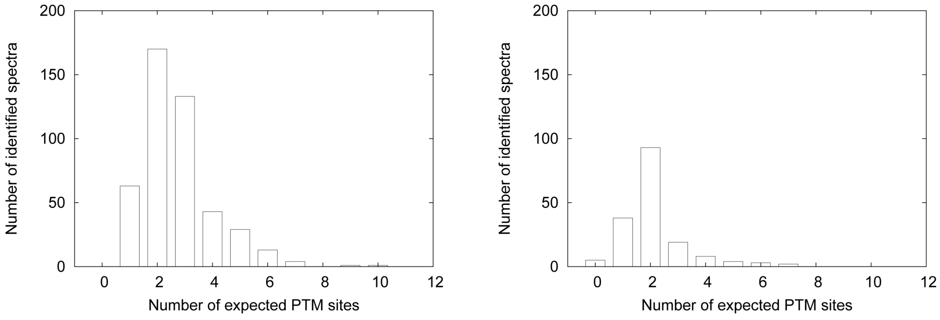


Figure 2. Spectral alignment

(a) A spectral alignment between the theoretical spectrum B = {0, 57, 144, 245, 302, 458, 559, 687} of a protein GSTGRTK and the theoretical spectrum B^* = {0, 57, 304, 375, 432, 588, 659, 787} of a modified protein GS[+160]T[-30]GRT[-30]K. The path from the top left corner (source) to the bottom right corner (sink) represents the alignment of B and B^* with three PTMs: +160 Da on the first S and −30 Da on the two T's. Diagonal and shift edges are shown in blue and red, respectively. The circles along the path denote the matching points in the alignment path. (b) A spectral alignment between a spectrum A = {0, 57, 375, 482, 588, 659, 787} generated from GS[+160]T[-30]GRT[-30]K and the theoretical spectrum B. Because mass 304 is missing in A, the PTM on the first S and the PTM on the first T are represented by a single shift edge (+130 Da) with a modification number 2. Another missing mass 432 in A results in replacing two consecutive diagonal edges by one diagonal edge. In addition, mass 482 is a noise mass. (c) A diagonal alignment between the spectrum A and the theoretical spectrum B (for a set of mass shifts S_1 = {−30, 160} and F = 3). The diagonal grid of A and B has 10 diagonal lines with offsets -90, -60, -30, 0, 100, 130, 160, 290, 320, and 480. The path from the source to the sink represents a diagonal alignment of spectrum A and protein B. The circles along the path denote diagonal points: blue ones have weight 1 and red ones have weight 0.



(a)

(b)

Figure 3. The histograms of numbers of expected PTM sites in 629 spectra identified from the histone H4 data set by MS-Align-E

(a) 457 spectra without unexpected PTM sites. (b) 172 spectra with a single unexpected PTM site.

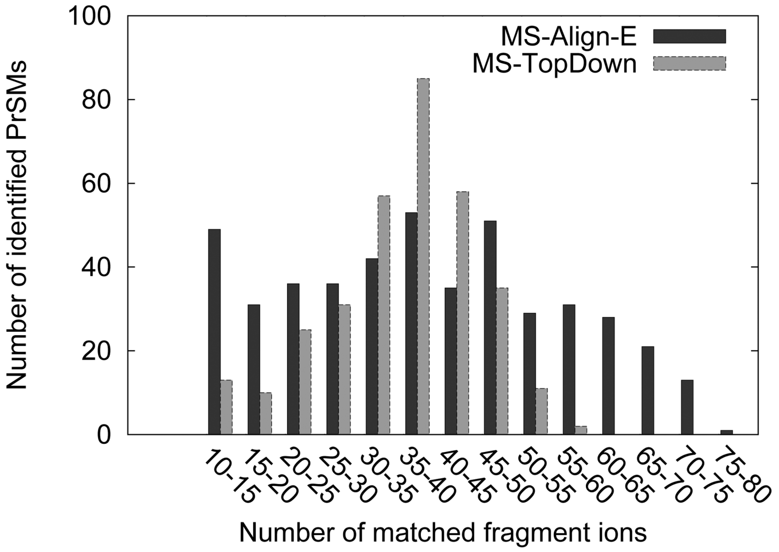
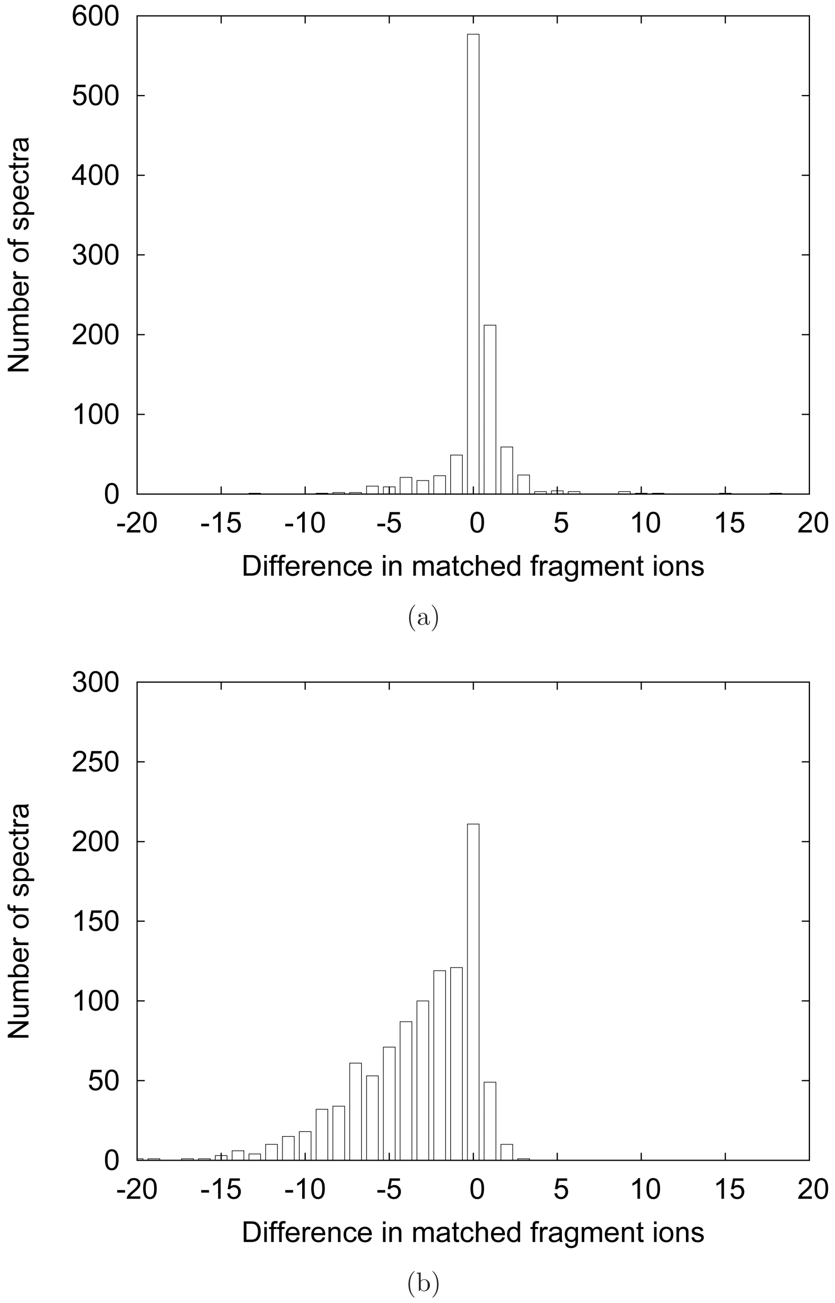


Figure 4. Comparison of MS-TopDown and MS-Align-E

MS-TopDown and MS-Align-E were applied to align 1, 626 ETD spectra with the histone H4 protein with one unexpected PTM site, and the numbers of identified PrSMs and matched fragment ions are reported.



(a)

(b)

Figure 5. Comparison of the numbers of matched fragment ions of the PrSMs identified in the histone H4 data set by MS-Align-E and ProSightPC

(a) The "absolute mass" mode. A total of 1024 spectra were identified by both MS-Align-E and ProSightPC. For each of the 1024 spectra, the numbers *M* and *P* of matched fragment ions of the PrSMs identified by MS-Align-E and ProSightPC are reported and the difference $M - P$ is computed. Of the 1024 spectra, MS-Align-E reported more matched fragment ions than ProSightPC for 312 spectra; ProSightPC reported a better proteoform than MS-Align-E for 135 spectra; and the two tools reported the same number of matched fragment ions for 577 spectra. (b) The "absolute mass" mode combined with the Δ*m* mode. A total of 1011 spectra were identified by both MS-Align-E and ProSightPC. Of the 1011 spectra, MS-

Align-E reported more matched fragment ions than ProSightPC for 60 spectra; ProSightPC reported more matched fragment ions than MS-Align-E for 740 spectra.

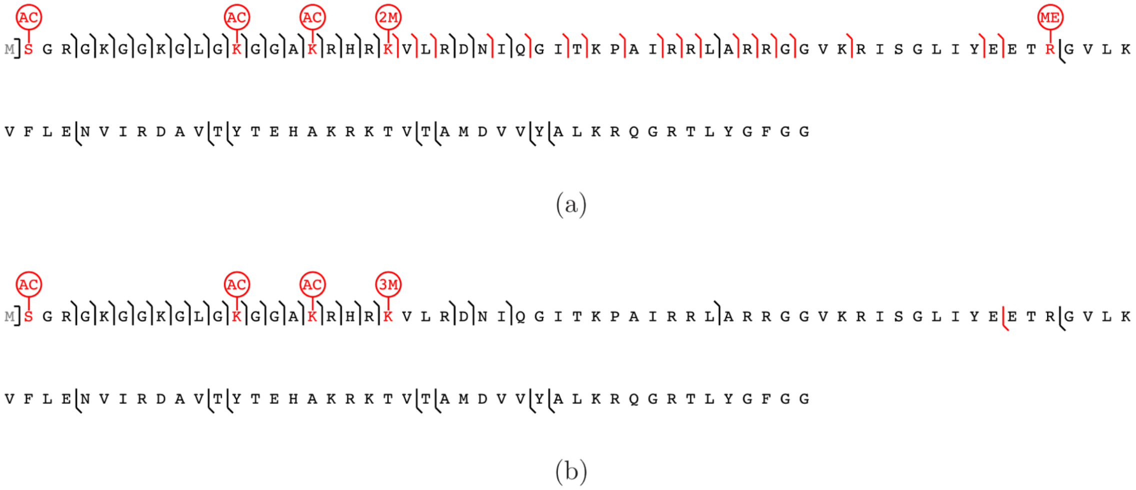

AC
AC
AC
2M
ME
M S G R G K G G K G L G K G G A K R H R K V L R D N I Q G I T K P A I R R L A R R G G V K R I S G L I Y E E T R G V L K
V F L E N V I R D A V T Y T E H A K R K T V T A M D V V Y A L K R Q G R T L Y G F G G
(a)
AC
AC
AC
3M
M S G R G K G G K G L G K G G A K R H R K V L R D N I Q G I T K P A I R R L A R R G G V K R I S G L I Y E E T R G V L K
V F L E N V I R D A V T Y T E H A K R K T V T A M D V V Y A L K R Q G R T L Y G F G G
(b)

Figure 6. MS-Align-E and the "absolute mass" mode of ProSightPC reported two different proteoforms for the spectrum of scan number 2, 062 in histone H4 spectral data set

(a) The proteoform reported by MS-Align-E has 47 matched fragment ions. (b) The proteoform reported by ProSightPC has 30 matched fragment ions. The ']' symbol right to the first methionine residue represents N-terminal methionine excision. Residues with PTMs are shown in red. AC, ME, 2M and 3M stand for acetylation, methylation, dimethylation, and trimethylation, respectively. Red lines represent matched fragment ions identified by only one tool; black lines represent matched fragment ions identified by two tools.

Table 1

Expected PTM types in the identification of proteoforms of histone H4.

PTM type	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