Published in final edited form as:
J Proteome Res. 2019 March 01; 18(3): 878–889. doi:10.1021/acs.jproteome.8b00562.

A Markov chain Monte Carlo method for estimating the statistical significance of proteoform identifications by top-down mass spectrometry

Qiang Kou[†], **Zhe Wang**[‡], **Rachele A. Lubeckyj**[¶], **Si Wu**[‡], **Liangliang Sun**[¶], and **Xiaowen Liu**[†,§]

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

[‡]Department of Chemistry and Biochemistry, The University of Oklahoma

[¶]Department of Chemistry, Michigan State University

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

Abstract

Top-down mass spectrometry is capable of identifying whole proteoform sequences with multiple post-translational modifications because it generates tandem mass spectra directly from intact proteoforms. Many software tools, such as ProSightPC, MSPathFinder, and TopMG, have been proposed for identifying proteoforms with modifications. In these tools, various methods are employed to estimate the statistical significance of identifications. However, most existing methods are designed for proteoform identifications without modifications, and the challenge remains for accurately estimating the statistical significance of proteoform identifications with modifications. Here we propose TopMCMC, a method that combines a Markov chain random walk algorithm and a greedy algorithm for assigning statistical significance to matches between spectra and protein sequences with variable modifications. Experimental results showed that TopMCMC achieved high accuracy in estimating *E*-values and false discovery rates of identifications in top-down mass spectrometry. Coupled with TopMG, TopMCMC identified more spectra than the generating function method from an MCF-7 top-down mass spectrometry data set.

Keywords

top-down mass spectrometry; statistical significance estimation; Markov chain Monte Carlo

Introduction

Top-down mass spectrometry (MS) has gained increasing attention in the past decade because of its ability to explore complex proteoforms.[1,2] By analyzing intact proteoforms

xwliu@iupui.edu, Phone: +1-317-278-7613. Fax: +1-317-278-9201.

Availability
The source code of TopMCMC is available at https://github.com/toppic-suite/toppic-suite.

Supporting information
The following Supporting Information is available free of charge on the ACS Publications website.

instead of short protein fragments,[3] top-down MS is capable of providing whole sequence information of proteoforms, many of which contain primary structural alterations, such as amino acid substitutions, post-translational modifications (PTMs), and terminal truncations.[4,5] Identification and characterization of these proteoforms aid researchers in answering questions in basic and translational research.[6,7]

Assigning accurate statistical significance to proteoform identifications is an important step in top-down mass spectral interpretation. [8,9] 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,[10] 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.[11–14] 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.[15]

The second is the *generating function* method,[15,16] 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. [17] Importance sampling methods, such as direct probability distribution (DPR), are often used in Monte Carlo simulation to estimate probabilities of extremely rare events.[18] Mohimani et al.[19] 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.[4,20] 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 paper, 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,[21] 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[9] and used in MS-Align+,[22] TopPIC,[23] and MS-PathFinder.[24]

In this paper, 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.[19] 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 method and the greedy algorithm, TopMCMC is capable of efficiently assigning protein-level statistical significance to PrSMs. We used two evaluation methods to test 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,[5] we identified more top-down mass spectra from an MCF-7 data set than TopMG with the generating function method.

Methods

Data sets

Four top-down MS data sets were used in this study: the first was generated from the human histone H3 protein, the second from the human histone H4 protein, the third from *Escherichia coli* (EC) K-12 MG1655 cells, and the fourth from human MCF-7 cells. The first three data sets were reported in previous studies,[23,25,26] and the fourth data set was generated for this study.

The histone H3 data set was generated by Tian et al.25 A total of 7.5 μg purified HeLa core histone proteins was first separated using a Jupiter C5 column (Phenomenex, Torrance, CA, USA). Two Cheminert column selector systems (VICI, Houston, TX, USA) were used to collect fractions of interest. Each collected fraction was further separated by a weak cation exchange hydrophilic interaction liquid chromatography (WCX-HILIC) system, which was coupled with an LTQ Orbitrap Velos mass spectrometer (Thermo Fisher Scientific, Waltham, MA). MS and MS/MS acquisitions were performed with a resolution of 60 000. The isolation window was 1.5 *m/z*, and alternating collision-induced dissociation (CID) and electron-transfer dissociation (ETD) fragmentation was performed to analyze precursor ions. Although all histone protein families were analyzed in Ref 25, we used only the data set of the histone H3 protein family, which contains 3 462 CID and 3 462 ETD top-down MS/MS spectra.

The histone H4 data set was reported in Ref. 26. Core histone proteins of primary normal human dermal fibroblasts (NHDFs) were purified using a histone purification kit (Active Motif, Carlsbad, CA). A total of 10 *μ*g core histone proteins were first separated by a reversed phase liquid chromatography (RPLC) column. The histone H4 protein was collected and further analyzed by a hydrophilic interaction liquid chromatography (HILIC) column coupled with an LTQ Orbitrap Velos mass spectrometer (Thermo Scientific, Waltham, MA). A high resolution 60 000 was used to acquire MS and MS/MS spectra. In each MS spectrum, the five most intense precursor ions were selected for MS/MS analysis. The precursor ions were collected with an isolation window of 3 *m/z* and analyzed by alternating CID and ETD fragmentation. The data set contains 1 626 CID and 1 626 ETD top-down MS/MS spectra.

The third data set was generated for studying proteoforms in EC K-12 MG1655 cells.[23] Intact proteins extracted from EC K-12 MG1655 cells were separated by an RPLC system coupled with an LTQ Orbitrap Velos mass spectrometry (Thermo Fisher Scientific, Waltham, MA). Alternating CID and ETD fragmentation was performed to analyze the 4 most intense precursor ions in each MS spectrum. The isolation window was 3 *m/z*. With a resolution of 60 000, a total of 2 027 CID and 2 027 ETD top-down MS/MS spectra were collected.

The fourth data set were generated from MCF-7 cells. 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 *μ*m/360 *μ*m 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.[27,28] 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, [29] 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 microscans was 4 and 3 for MS and MS/MS, respectively. A total of 1 523 top-down higher-energy collisional dissociation (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).[30] The mzML files were further deconvoluted by TopFD (version 1.1.2, http://proteomics.informatics.iupui.edu/software/toppic/), which converted all MS/MS spectra into lists of neutral fragment masses.

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[24,31,32] 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 PrecMass(S) − Mass(H_2O), where PrecMass(S) is the monoisotopic neutral precursor mass of S and Mass(H_2O) 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, $F = a_1a_2 \dots a_n$, 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. [26] 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 $1 \le i \le n - 1$, 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 $t(F) = \{f_1, \dots, f_{n-1}, g_i, \dots, g_{n-1}\}$. 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 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.

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 “Sequences of standard amino acids”, 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, **147**, 114, 156, 129, **167**, 128 is a modified proteoform of protein 57, **131**, 114, 156, 129, **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 PScore(S, P, V). That is, PScore(S, P, V) = $\max_{F \in D(P, V)}$ 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 PScore(S, P, V) is zero.

In this paper, we study protein-level statistical significance of matches between proteins and spectra. When 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 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.

Similarity scores between spectra and proteins are computed in two phases in database search. In the first phase, tens of database proteins have been reported for a query spectrum by a filtering algorithm, and we need to compute the similarity score between the spectrum and each of the database proteins. In the second phase, a PrSM with a similarity score has been identified, and the MCMC method is used to estimate the statistical significance of the identification. In the MCMC method, ten of thousands of random proteins are simulated, and we need to compute the P-score between the spectrum and each of the random proteins. The number of similarity score computations for a query spectrum in the second phase is usually more than 1 000 times of that in the first phase. A dynamic programming method is often used to compute similarity scores in the first phase, but it is slow for the second phase.

To address the problem, we propose a greedy algorithm (Fig. 2) to quickly estimate PScore(S, P, V). We first define neighbor proteoforms used 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 of one PTM in F_1 and *vice versa*. For example, F = 57, **147**, 114, 156, 129, **167**, 128 is a proteoform of protein 57, 131, 114, 156, 129, 87, 128 with two PTMs {16, 80}, and F' = 57, 131, 114, **172**, 129, **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 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 PScore(S, P, V).

Representing proteins by Markov chains

Similar to the method proposed by Mohimani et al.,[19] 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 ($1 \le i \le n$) in two proteins $a_1a_2 \ldots a_n$ and $b_1b_2 \ldots b_n$ are a matched ass 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, **71**, 114, 156, **129**, 57, 128 and 57, **87**, 114, 156, **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 $P = a_1a_2 \ldots a_n$ with a residue mass $m = \sum_{i=1}^{n} a_i$. Let $P' = b_1b_2 \ldots b_n$ 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 $\sum_{i=1}^{n}\left(a_i - 1\right)(n - 1) = (m - n)(n - 1)$. 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 (Fig. 1). 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, [33] 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 $\Omega_{n,m}$.

The direct probability redistribution method

Let X be a random variable for the similarity score PScore(S, P, V) between a spectrum S and a random protein $P \in \Omega_{n,m}$ with a fixed multiset $V = \{v_1, v_2, \ldots, v_k\}$ of PTMs. The space of X is $\{0, 1, \ldots, m\}$, where m is the number of masses in the spectrum S. When the spectrum S and multiset V are fixed, the score 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.

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.[18] 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 (Fig. 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 $\frac{z_i/u_i}{\sum_{k=0}^{m} z_i/u_k}$. More details of the DPR method can be found in Ref. 18.

The oversampling factors $u_0, \ldots, u_m$ are important parameters for accurate estimation of rare event probabilities. Haraszti et al. [18] 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 (Fig. 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.

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 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 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 PScore(S, P, V) ≥ 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 $p(i, j, t, V) = \Pr(\mathrm{PScore}(S, P, V) \geq t)$. Let $X(t, V)$ be a random variable representing the number of proteins in D with a score PScore(S, P, V) ≥ 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 PScore(S, P, V) ≥ 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 $\max_{P \in D, V \in \Phi_k}$ PScore(S, P, V) ≥ 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 PScore(S, P, V) ≥ 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. (See the supplementary material for details).

To speed up the computation, the greedy algorithm in Fig. 2 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 PScore(S, P*, V). Second, we use the DPR method to compute the probability that the estimation of 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 $p(i, j, t, V)$.

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 $A_5 = B_5$. 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 Fig. 5.

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 z 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 z is narrowly concentrated and has a small relative standard deviation (Fig. S1 in the supplementary material).

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 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 $x + a > 0$.

Results

The proposed TopMCMC 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.

Evaluation of the greedy algorithm

The greedy algorithm in Fig. 2 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).[34] Acetylation, methylation, dimethylation, trimethylation, and phosphorylation were considered as variable PTMs (Table S1 in the supplementary material). 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 reported by the greedy algorithm is an estimation of PScore(S, P, V). 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 Fig. 2), and the best score was reported. The parameter settings of TopMG can be found in Table S2 in the supplementary material.

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 (Table S3 in the supplementary material). 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. Fig. 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. In the MCMC method, a large variance in the difference of the two scores significantly affects the accuracy of estimated p-values, but a large average difference does not. For example, when the variance of the differences is 0, for a match with a G-score g and a P-score p, the probability that a random protein has a G-score $\geq g$ is the same as the probability that a random protein has a P-score $\geq p$. The reason is that the difference between the G-score and P-score of a random protein is fixed. 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. The greedy method introduces more errors for matches with >5 PTMs than those with ≤ 5 PTMs.

Evaluation based on *p*-values

The bipartite database strategy[35] 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.

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[36] demonstrated that *P. furiosus* proteins are a good choice for entrapment sequences because they have a long evolutionary distance with human sequences,

TopMG[5] 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 (Table S1 in the supplementary material). The error tolerance for precursor and fragment masses was set to 15 ppm, at most 5 variable PTMs were allowed in an identified proteoform, 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. The parameter settings of TopMG and TopMCMC are given in Table S4 and S5 in the supplementary material. Of the 6 824 spectra, 2 638 were matched to proteoforms of the entrapment sequences (Table S6 in the supplementary material).

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} (Fig. 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).[35] 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 (Fig. S2 and S3 in supplementary material).

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] (Fig. 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. Fig. 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.

Evaluation based on FDRs

We also evaluated the accuracy of TopMCMC using a false discovery rate (FDR)-based method.[9,37] 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),[38] and the other by the eTDA estimator.[37] 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. A brief description of the eTDA method is given in the supplementary material. 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 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. Many spectra in the EC data set were generated from proteoforms without modifications, and it was unnecessary to use TopMG to analyze these spectra. In the first step, all mass spectra in the EC data set were searched against the EC proteome database using TopPIC[23] to quickly identify PrSMs without modifications (some may contain terminal truncations), speeding up the following TopMG analysis. In addition, mass shifts identified by TopPIC were used to find common PTMs in the data set. In TopPIC, one unexpected mass shift was allowed in a proteoform, and other parameter settings can be found in Table S7 in the supplementary material. With a 1% spectrum level FDR, a total of 1 920 PrSMs from 178 proteins were identified, including 470 PrSMs with unexpected mass shifts (Table S8 in the supplementary material). Many mass shifts in the 470 PrSMs can be explained by common PTMs (Table 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, the 1 450 spectra matched to proteoforms without mass shifts in the previous step were excluded, and the remaining 2 604 spectra (including the 470 spectra identified with mass shifts in the 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. Other parameter settings of TopMG can be found in Table S9 in the supplementary material.

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} (Fig. 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.

Discriminative capacity

We compared the TopMCMC method and the generating function approach[9,16] 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[34] 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 115 proteins, including 400 PrSMs without unexpected mass shifts (Table S11 in the supplementary material). 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 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.

With a 5% spectrum level FDR, TopMG+MCMC and TopMG+GF identified 161 and 133 PrSMs, respectively (Fig. 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. Four proteoforms missed by TopMG+GF are given in Fig. S4-S7 in the supplementary material. 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.

Discussion

There are two main differences between TopMCMC and MS-DPR[19] 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. In MS-DPR, 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 MS-DPR 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 score between the query spectrum and the best candidate proteoform of the protein. Because the number of all candidate proteoforms grows exponentially with the number of PTM sites, it is inefficient to compute the score by enumerating all candidate proteoforms. Although the mass graph alignment algorithm in TopMG solves the problem in a polynomial time complexity when the number of PTMs in a proteoform is limited, its running time is still unacceptable when a spectrum is aligned with tens of thousands of proteins. To address this problem, a greedy method is used in TopMCMC to speed up the computation.

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,[39] 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 (Fig. S2 and S3 in supplementary material) 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.

The TopMCMC method still has some limitations. First, a greedy algorithm was introduced to speed up the estimation of the similarity score. The greedy algorithm often fails to report accurate scores for PrSMs with a large number of variable PTMs, and affects the accuracy of estimated *p*-values and *E*-values. TopMCMC may introduce large errors to estimated *p*-values and *E*-values of PrSMs with more than 5 PTMs. A fast and accurate method for computing similarity scores can further improve the performance of TopMCMC. Second, a simple shared mass counting score is used, and the peak intensity information is ignored.

Using a scoring function that takes account peak intensities into consideration can further improve the discriminative capacity of TopMCMC.

Conclusions

In this paper, we proposed a new MCMC-based method for estimating the statistical significance for proteoform identifications with variable PTMs in top-down MS. The experiments showed that TopMCMC achieved high accuracy in estimating *p*-values and *E*-values and outperformed the generating function method in distinguishing correct identifications from incorrect ones.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

The research was supported by the National Institute of General Medical Sciences, National Institutes of Health (NIH) through Grants R01GM118470 and R01GM125991.

References

(1). Cui W; Rohrs HW; Gross ML Top-down mass spectrometry: recent developments, applications and perspectives. Analyst 2011, 136, 3854–3864. [PubMed: 21826297]

(2). Aebersold R et al. How many human proteoforms are there? Nature Chemical Biology 2018, 14, 206–214. [PubMed: 29443976]

(3). Catherman AD; Skinner OS; Kelleher NL Top Down proteomics: facts and perspectives. Biochemical and Biophysical Research Communications 2014, 445, 683–93. [PubMed: 24556311]

(4). Young NL; DiMaggio PA; Plazas-Mayorca MD; Baliban RC; Floudas CA; Garcia BA High throughput characterization of combinatorial histone codes. Molecular & Cellular Proteomics 2009, 8, 2266–2284. [PubMed: 19654425]

(5). Kou Q; Wu S; Tolić N; Paša-Tolić L; Liu Y; Liu X A mass graph-based approach for the identification of modified proteoforms using top-down tandem mass spectra. Bioinformatics 2016, 33, 1309–1316.

(6). Peleg S et al. Altered histone acetylation is associated with age-dependent memory impairment in mice. Science 2010, 328, 753–756. [PubMed: 20448184]

(7). Dong X; Sumandea CA; Chen Y-C; Garcia-Cazarin ML; Zhang J; Balke CW; Sumandea MP; Ge Y Augmented phosphorylation of cardiac troponin I in hypertensive heart failure. Journal of Biological Chemistry 2012, 287, 848–857. [PubMed: 22052912]

(8). LeDuc RD; Taylor GK; Kim Y-B; Januszyk TE; Bynum LH; Sola JV; Garavelli JS; Kelleher NL ProSight PTM: an integrated environment for protein identification and characterization by top-down mass spectrometry. Nucleic Acids Research 2004, 32, W340–W345. [PubMed: 15215407]

(9). Liu X; Segar MW; Li SC; Kim S Spectral probabilities of top-down tandem mass spectra. BMC Genomics 2014, 15, S9.

(10). Noble WS; MacCoss MJ Computational and statistical analysis of protein mass spectrometry data. PLoS Computational Biology 2012, 8, e1002296. [PubMed: 22291580]

(11). Keller A; Nesvizhskii AI; Kolker E; Aebersold R Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Analytical Chemistry 2002, 74, 5383–5392. [PubMed: 12403597]

(12). Sadygov RG; Yates JR A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. Analytical Chemistry 2003, 75, 3792–3798. [PubMed: 14572045]

(13). Sadygov RG; Liu H; Yates JR Statistical models for protein validation using tandem mass spectral data and protein amino acid sequence databases. Analytical Chemistry 2004, 76, 1664–1671. [PubMed: 15018565]

(14). Nesvizhskii AI; Aebersold R Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem MS. Drug Discovery Today 2004, 9, 173–181. [PubMed: 14960397]

(15). Kim S; Mischerikow N; Bandeira N; Navarro JD; Wich L; Mohammed S; Heck AJ; Pevzner PA The generating function of CID, ETD, and CID/ETD pairs of tandem mass spectra: applications to database search. Molecular & Cellular Proteomics 2010, 9, 2840–2852. [PubMed: 20829449]

(16). 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]

(17). Rubinstein RY; Kroese DP Simulation and the Monte Carlo method, 2nd ed.; John Wiley & Sons, Inc: Hoboken, New Jersey, 2008.

(18). Haraszti Z; Townsend JK The theory of direct probability redistribution and its application to rare event simulation. ACM Transactions on Modeling and Computer Simulation (TOMACS) 1999, 9, 105–140.

(19). Mohimani H; Kim S; Pevzner PA A new approach to evaluating statistical significance of spectral identifications. Journal of Proteome Research 2013, 12, 1560–1568. [PubMed: 23343606]

(20). Qiang K; Si W; Xiaowen L Systematic evaluation of protein sequence filtering algorithms for proteoform identification using top-down mass spectrometry. Proteomics 2018, 18, 1700306.

(21). Karabacak NM; Li L; Tiwari A; Hayward LJ; Hong P; Easterling ML; Agar JN Sensitive and specific identification of wild type and variant proteins from 8 to 669 kDa using top-down mass spectrometry. Molecular & Cellular Proteomics 2009, 8, 846–856. [PubMed: 19074999]

(22). 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, 11, M111.008524.

(23). Kou Q; Xun L; Liu X TopPIC: a software tool for top-down mass spectrometry-based proteoform identification and characterization. Bioinformatics 2016, 32, 3495–3497. [PubMed: 27423895]

(24). Park J et al. Informed-Proteomics: open-source software package for top-down proteomics. Nature Methods 2017, 14, 909–914. [PubMed: 28783154]

(25). Tian Z; Tolić N; Zhao R; Moore RJ; Hengel SM; Robinson EW; Stenoien DL; Wu S; Smith RD; Paša-Tolić L Enhanced top-down characterization of histone post-translational modifications. Genome Biology 2012, 13, R86. [PubMed: 23034525]

(26). Liu X; Hengel S; Wu S; Tolić N; Paša-Tolić L; Pevzner PA Identification of ultramodified proteins using top-down tandem mass spectra. Journal of Proteome Research 2013, 12, 5830–5838. [PubMed: 24188097]

(27). Wojcik R; Dada OO; Sadilek M; Dovichi NJ Simplified capillary electrophoresis nanospray sheath-flow interface for high efficiency and sensitive peptide analysis. Rapid Communications in Mass Spectrometry 2010, 24, 2554–2560. [PubMed: 20740530]

(28). Sun L; Zhu G; Zhao Y; Yan X; Mou S; Dovichi NJ Ultrasensitive and fast bottom-up analysis of femtogram amounts of complex proteome digests. Angewandte Chemie International Edition 2013, 52, 13661–13664. [PubMed: 24173663]

(29). Lubeckyj RA; McCool EN; Shen X; Kou Q; Liu X; Sun L 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, 89, 12059–12067. [PubMed: 29064224]

(30). Kessner D; Chambers M; Burke R; Agus D; Mallick P ProteoWizard: open source software for rapid proteomics tools development. Bioinformatics 2008, 24, 2534–6. [PubMed: 18606607]

(31). 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, 320–332. [PubMed: 10757168]

(32). Liu X; Inbar Y; Dorrestein PC; Wynne C; Edwards N; Souda P; White-legge 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]

(33). Motwani R; Raghavan P Randomized Algorithms; Cambridge University Press: New York, NY, USA, 1995.

(34). The UniProt Consortium, UniProt: a hub for protein information. Nucleic Acids Research 2015, 43, D204–D212. [PubMed: 25348405]

(35). Granholm V; Noble WS; Käll L 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 2011, 10, 2671–2678. [PubMed: 21391616]

(36). Vaudel M; Burkhart JM; Breiter D; Zahedi RP; Sickmann A; Martens L A complex standard for protein identification, designed by evolution. Journal of Proteome Research 2012, 11, 5065–5071. [PubMed: 22489649]

(37). Gupta N; Bandeira N; Keich U; Pevzner PA Target-decoy approach and false discovery rate: when things may go wrong. Journal of the American Society for Mass Spectrometry 2011, 22, 1111–1120. [PubMed: 21953092]

(38). Elias JE; Gygi SP Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nature methods 2007, 4, 207. [PubMed: 17327847]

(39). Kou Q; Zhu B; Wu S; Ansong C; Tolić N; Paša-Tolić L; Liu X Characterization of Proteoforms with Unknown Post-translational Modifications Using the MIScore. Journal of Proteome Research 2016, 15, 2422–2432. [PubMed: 27291504]

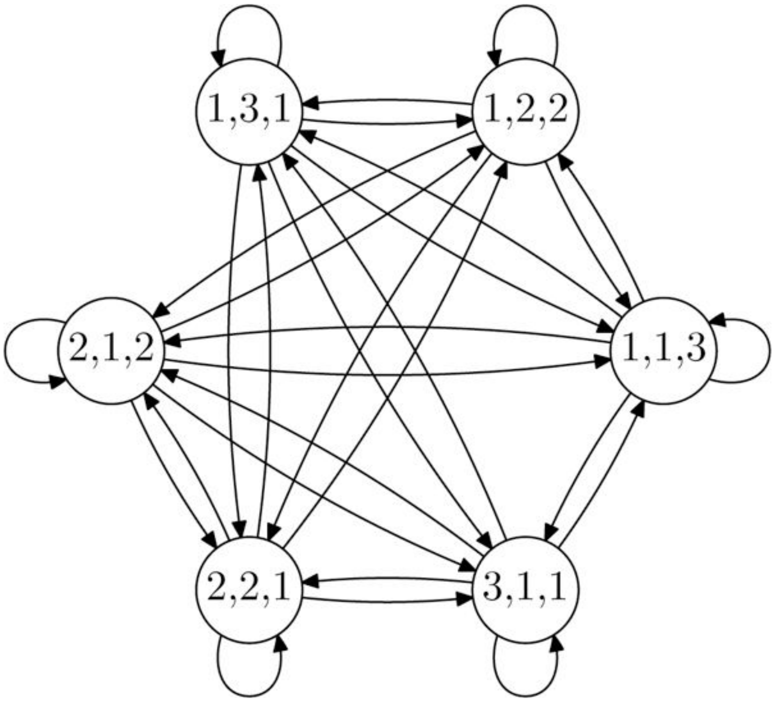
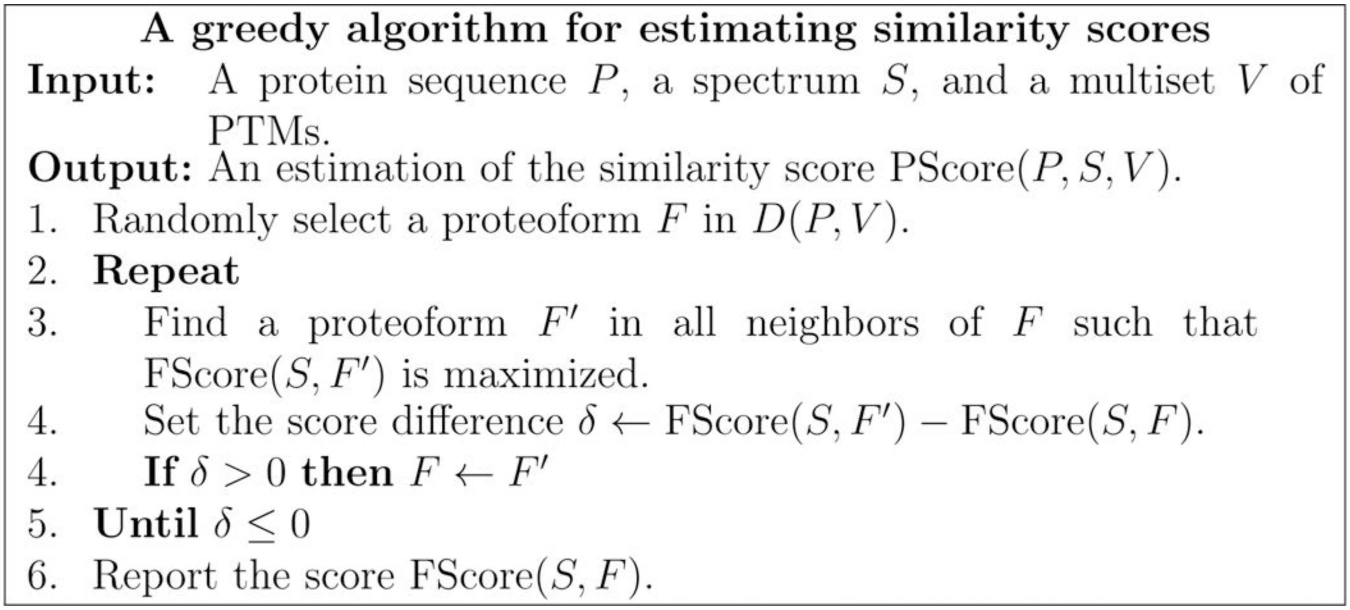


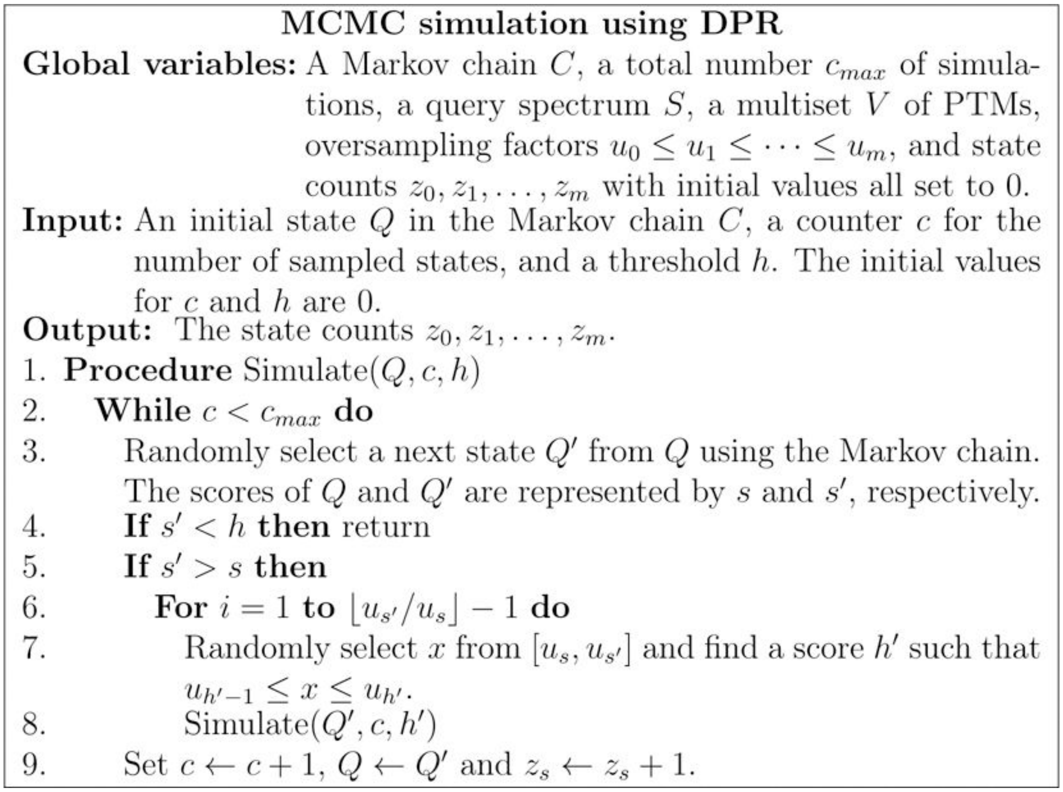
Figure 1:

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 $\frac{1}{5}$.



```
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 PScore$(P, S, V)$.
Randomly select a proteoform $F$ in $D(P, V)$.
Repeat
    Find a proteoform $F'$ in all neighbors of $F$ such that FScore$(S, F')$ is maximized.
    Set the score difference $\delta \leftarrow$ FScore$(S, F')$ $-$ FScore$(S, F)$.
    If $\delta > 0$ then $F \leftarrow F'$
Until $\delta \leq 0$
Report the score FScore$(S, F)$.
```

Figure 2:
A greedy algorithm for estimating similarity scores.



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.

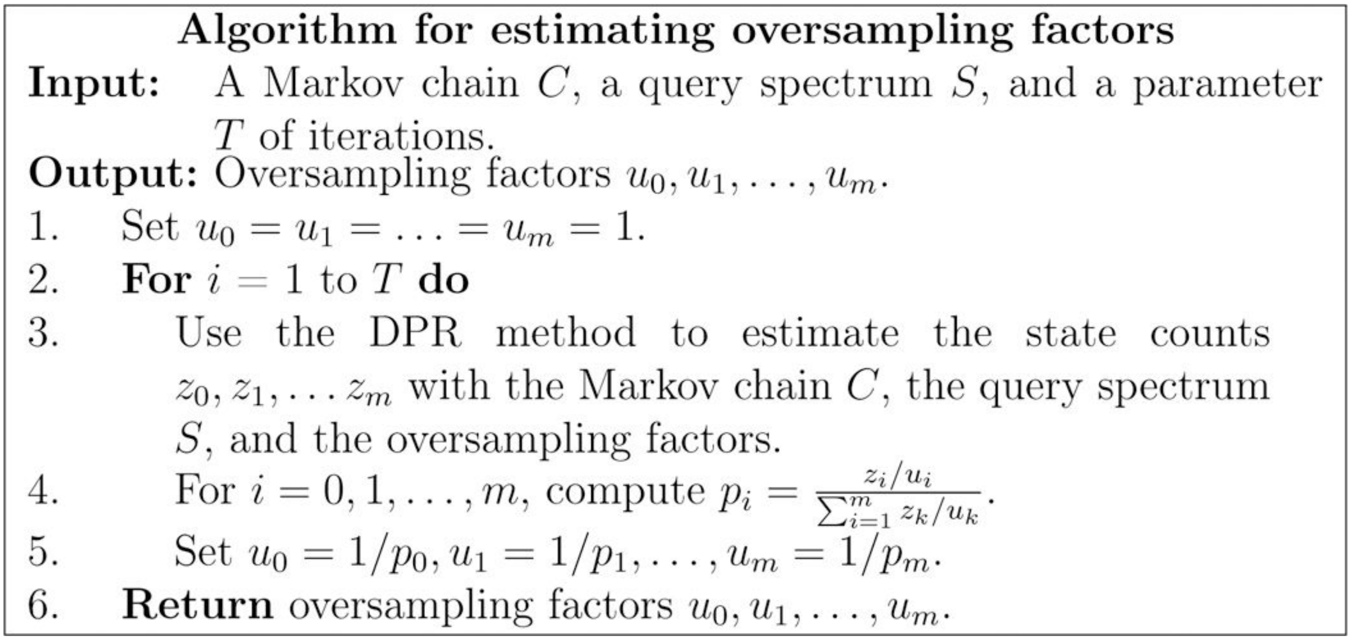
Input: An initial state Q in the Markov chain C, a counter c for the number of sampled states, and a threshold h. The initial values for c and h are 0.

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

```
Procedure Simulate($Q, c, h$)
  While $c < c_{max}$ do
    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.
    If $s' < h$ then return
    If $s' > s$ then
      For $i = 1$ to $\lfloor u_{s'}/u_s \rfloor - 1$ do
        Randomly select $x$ from $[u_s, u_{s'}]$ and find a score $h'$ such that
        $u_{h'-1} \leq x \leq u_{h'}$.
        Simulate($Q', c, h'$)
    Set $c \leftarrow c + 1$, $Q \leftarrow Q'$ and $z_s \leftarrow z_s + 1$.
```

Figure 3:

MCMC simulation using DPR.



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 4:

The algorithm for estimating oversampling factors.

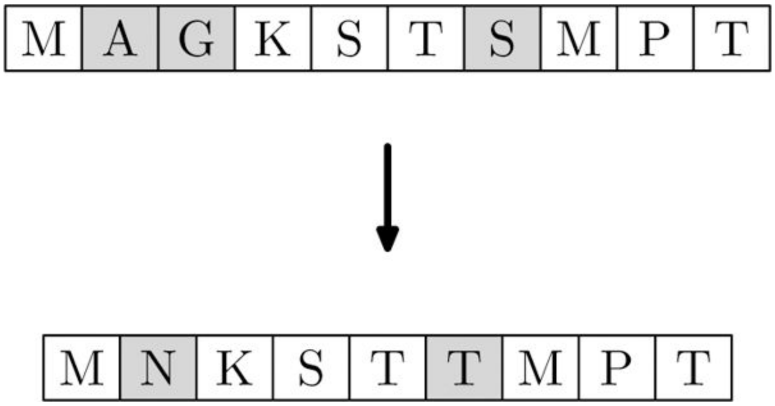


Figure 5:

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.

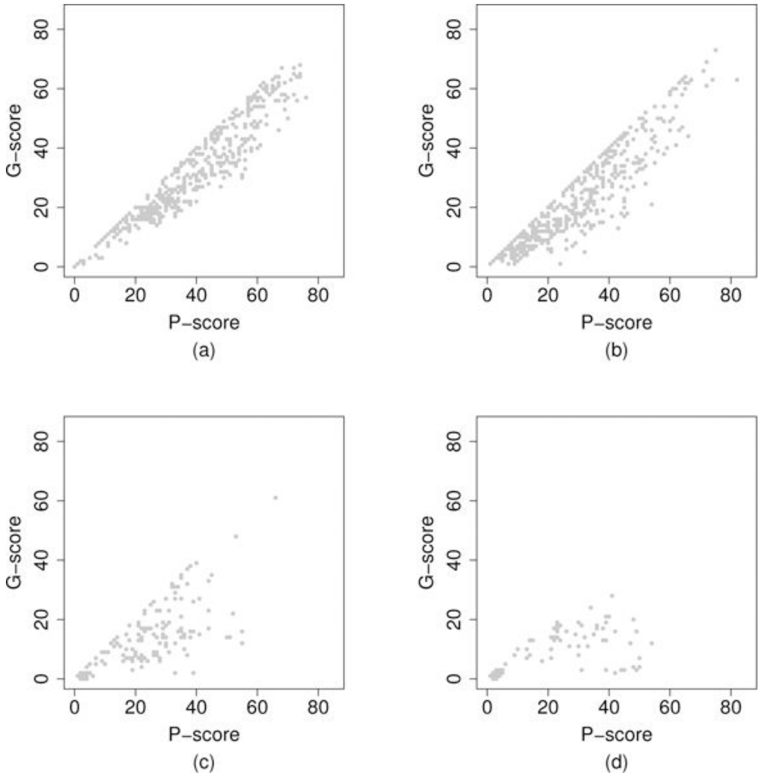


Figure 6:
Scatter plots of the P-scores and G-scores of the 1 112 protein spectrum matches in the histone H4 data set with various numbers of PTMs: (a) 0 − 2 PTMs; (b) 3 − 5 PTMs; (c) 6 − 8 PTMs; (d) 9 − 10 PTMs.

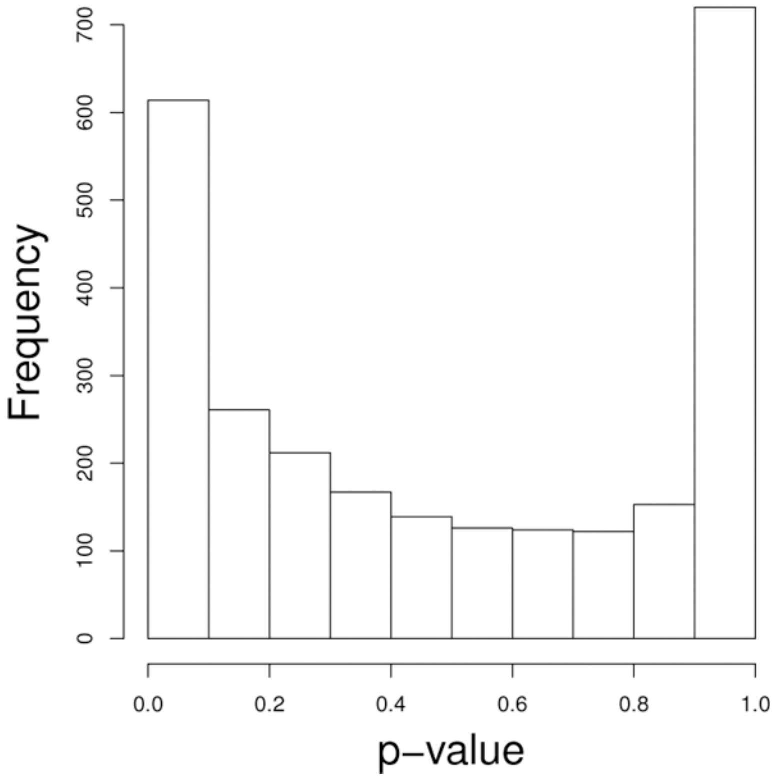


Figure 7:
The histogram of *p*-values reported by TopMCMC for the 2 638 entrapment PrSMs reported from the histone H3 data set. The *D* value (Kolmogorov-Smirnov statistic) between the empirical distribution of the *p*-values and the uniform distribution over [0, 1] is 0.1874.

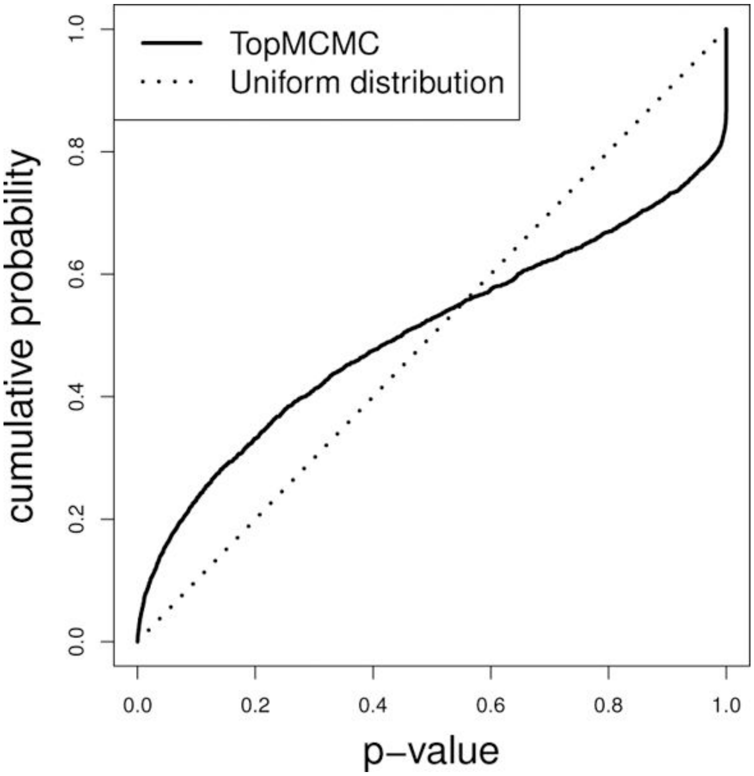


Figure 8:

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.

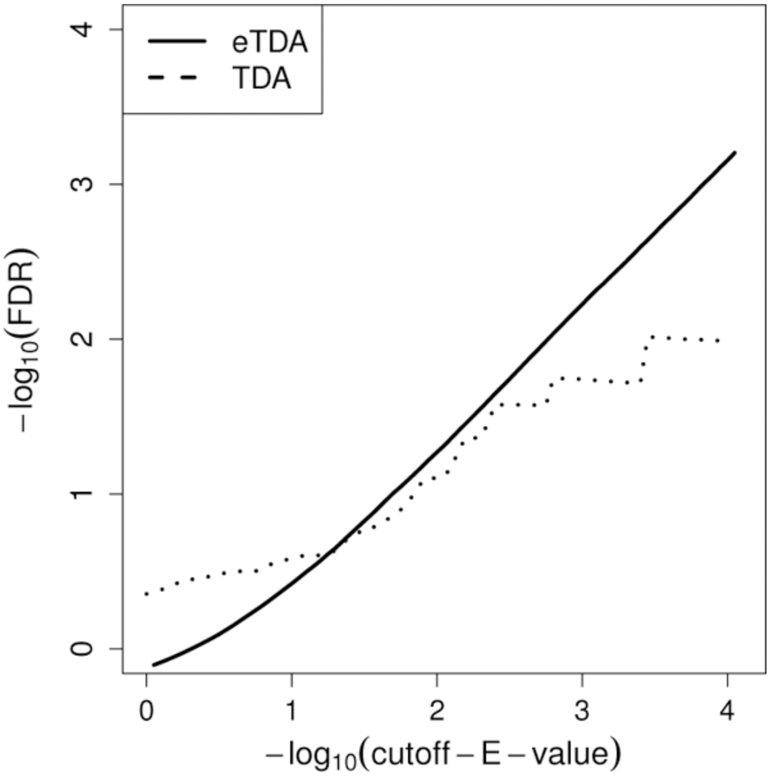


Figure 9:

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

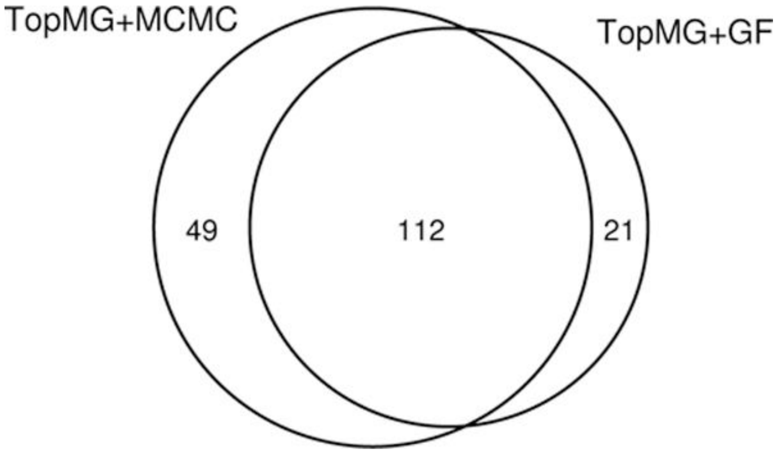


Figure 10:

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.

Table 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

Table 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