Zhu and Liu *BMC Bioinformatics* 2018, **19**(Suppl 9):280
https://doi.org/10.1186/s12859-018-2273-4

BMC Bioinformatics

RESEARCH Open Access

# A graph-based approach for proteoform identification and quantification using top-down homogeneous multiplexed tandem mass spectra

Kaiyuan Zhu[1] and Xiaowen Liu[2,3*]

*From* 13th International Symposium on Bioinformatics Research and Applications (ISBRA 2017)
Honolulu, Hawaii, USA. 30 May - 2 June 2017

## Abstract

**Background:** Top-down homogeneous multiplexed tandem mass (HomMTM) spectra are generated from modified proteoforms of the same protein with different post-translational modification patterns. They are frequently observed in the analysis of ultramodified proteins, some proteoforms of which have similar molecular weights and cannot be well separated by liquid chromatography in mass spectrometry analysis.

**Results:** We formulate the top-down HomMTM spectral identification problem as the minimum error *k*-splittable flow problem on graphs and propose a graph-based algorithm for the identification and quantification of proteoforms using top-down HomMTM spectra.

**Conclusions:** Experiments on a top-down mass spectrometry data set of the histone H4 protein showed that the proposed method identified many proteoform pairs that better explain the query spectra than single proteoforms.

**Keywords:** Mass spectrometry, Top-down, Multiplexed mass spectra, Graph algorithms

## Background

In top-down mass spectrometry (MS), separating similar proteoforms is a challenging problem. A ultramodified protein may have many similar proteoforms with similar weights and different post-translational modification (PTM) patterns. These proteoforms are often not well separated in top-down MS analysis [1]. A *multiplexed tandem mass (MTM) spectrum* is generated when tandem mass spectrometry (MS/MS) is used to analyze two or more proteoforms with similar molecular masses that are not separated by protein separation methods [2]. Despite the complexity of MTM spectra, they have been extensively studied because the interpretation of these spectra provides valuable information about modifications and quantification of proteoforms of ultramodified proteins [1–3]. For example, MTM spectra are frequently observed and analyzed in studies of histone proteins, which play important roles in epigenetics and gene regulation [4, 5].

MTM spectra can be divided into two main types: *heterogeneous* multiplexed tandem mass (HetMTM) spectra and *homogeneous* multiplexed tandem mass (HomMTM) spectra. While HetMTM spectra are generated from proteoforms of two or more different proteins, HomMTM ones from proteoforms of the same protein with different PTM patterns. In data-independent acquisition MS, which has been rapidly developed in the past several years, the precursor ions in a large mass-to-charge ratio ($m/z$ value) interval are collected for MS/MS analysis, resulting

*Correspondence: xwliu@iupui.edu
[2]Department of BioHealth Informatics, Indiana University-Purdue University Indianapolis, 719 Indiana Avenue, 46202 Indianapolis, IN, USA
[3]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, 410 W. 10th Street, 46202 Indianapolis, IN, USA
Full list of author information is available at the end of the article

in complex HetMTM spectra [6, 7]. In spectral identification, a HetMTM spectrum is searched against a protein database to find $k$ proteins/peptides that best explain the spectrum [2], where $k$ is a user-defined parameter. The problem is computational challenging because its search space is proportional to $n^k$, where $n$ is the number of proteins/peptides in the database.

In the analysis of HomMTM spectra, we often focus on purified proteins, whose sequences are known. Let $P$ be an unmodified protein sequence and $S$ a HomMTM spectrum generated from $k$ modified proteoforms of $P$. Denote $\mathcal{Q}_M$ as the set of modified proteoforms of $P$ that match the precursor mass of $S$. The *HomMTM spectral identification problem* is to find $k$ proteoforms in $\mathcal{Q}_M$ and their relative abundances to maximize the similarity between the theoretical spectra of the proteoforms and the spectrum $S$ [1].

DiMaggio et al. first studied the HomMTM spectral identification problem and proposed a mixed integer linear optimization framework for solving it [1]. The proposed framework demonstrated good performance on the analysis of middle-down MTM spectra of histone proteins, but the exponential time complexity of integer linear optimization makes it inefficient for analyzing top-down HomMTM spectra of long protein sequences.

In top-down MS, many software tools have been developed for the identification of proteoforms with PTMs and other alterations [8–12]. However, these software tools are designed for analyzing tandem mass spectra from single proteoforms, not multiplexed ones. Using these tools to analyze an MTM spectrum reports only one proteoform instead of multiple ones.

We formulate the minimum error $k$-splittable flow (ME$k$SF) problem on graphs and convert the HomTMT spectral identification problem to the ME$k$SF problem. To our best knowledge, the ME$k$SF problem has not been studied. However, the maximum $k$-splittable flow (M$k$SF) problem, which is related to the ME$k$SF problem, has been extensively studied and has various applications in commodity transportation and telecommunication network optimization [13–18].

Let $G$ be a connected graph with edge capacities, a source vertex, and a sink vertex. A flow is $k$-splittable if it can be decomposed to $k$ or less than $k$ paths. These paths are neither required to be different, nor edge/vertex disjoint. The M$k$SF problem aims at finding a $k$-splittable flow in $G$ from the source to the sink such that the edge capacity constraints are not violated and the flow value is maximized.

Baier et al. [13, 14] first investigated the M$k$SF problem and proved the NP-hardness of the problem on directed graphs for $k = 2$. They proposed approximation algorithms with a performance ratio $\frac{2}{3}$ for the maximum 2 and 3-splittable flow problem and presented a $\frac{1}{2}$-approximation algorithm for the general M$k$SF problem. Koch and Spenke [15] studied the complexity and approximability of the M$k$SF problem for different values of $k \geq 2$ on directed and undirected graphs. In particular, they proved that the problem is NP-hard for $k = 2$ on directed and undirected graphs and showed that, for an arbitrary constant $k$, the problem cannot be approximated with a performance ratio better than $\frac{5}{6}$. Koch, Skutella and Spenke [16] decoupled the M$k$SF problem into two steps: the first step called *packing* finds the flow values of the $k$ paths in an optimal solution; the second step called *routing* reports the optimal paths of the $k$ flow values. The packing procedure was described for general directed graphs, while the routing for graphs with bounded treewidth. Finally, they proposed a polynomial algorithm for the M$k$SF problem on graphs of bounded treewidth when $k$ is a constant and presented a polynomial-time approximation scheme when $k$ is part of the input.

Unlike the M$k$SF problem, an instance graph of the ME$k$SF problem has capacities on vertices instead of edges. In addition, it is allowed that a flow violates the vertex capacity constrains. That is, the flow value on a vertex may be larger than its capacity. The difference between the flow value and the capacity on a vertex (the flow value may be smaller or larger than the capacity) is defined as the error of the vertex. Let $G$ be a connected graph with integer vertex capacities, a source vertex, and a sink vertex. Given a total integer flow value $f$, the objective of the ME$k$SF problem is to find a $k$-splittable flow $F$ in $G$ from the source to the sink such that the flow value of $F$ is $f$ and the sum of the errors on the vertices is minimized.

We prove that the ME$k$SF problem is NP-hard when $k$ is part of the input and propose a polynomial time algorithm for the problem on layered directed graphs when $k = 2$. We tested the algorithm on a top-down MS/MS data set of the human histone H4 protein. Experimental results showed that the proposed method identified many proteoform pairs (path pairs in the graph) that provided better explanation for the query spectra than single proteoforms reported by MS-Align-E [9], an existing tool for the identification of ultramodified proteins.

## Methods

### The ME$k$SF problem

Let $G = (V, E)$ be a directed graph with a source vertex $s$ and a sink vertex $t$. Each vertex $v \in V$ has a positive integer capacity $c(v) \in \mathbb{Z}^+$. Let $\mathcal{A}$ denote the set of all simple $s$-$t$-paths (without circles) in $G$. A $k$-splittable $s$-$t$-flow $F$ contains $k$ pairs $(A_1, f_1), \cdots, (A_k, f_k)$ where $A_i$ is a path in $\mathcal{A}$ and $f_i \in \mathbb{Z}^+$ is the integer flow value on $A_i$, for $1 \leq i \leq k$. The paths $A_1, \cdots, A_k$ may share vertices and/or edges. The flow value of $F$ is the sum

$\sum_{i=1}^{k} f_i$. Let $F(v)$ be the set of the pairs $(A_i, f_i)$ in $F$ satisfying that $A_i$ contains vertex $v \in V$. The flow value of $v$ is the sum of the flow values of the pairs in $F(v)$, denoted by $f(v) = \sum_{(A_i,f_i)\in F(v)} f_i$. The error on $v$ is the difference between the flow value and the capacity of the vertex, denoted by $\varepsilon(v) = |f(v) - c(v)|$. The error of $F$ is the sum or the errors of all vertices in $G$, denoted by $\varepsilon(F) = \sum_{v\in V} \varepsilon(v)$. The ME$k$SF problem is defined as follows.

**Definition 1** *Given a directed graph G with integer vertex capacities, a source vertex s and a sink vertex t, and an integer flow value f, the MEkSF problem is to find a k-splittable flow F in G from s to t such that the flow value of F is f and the error of F is minimized.*

#### The HomMTM spectral identification problem

When a purified protein is analyzed and the target protein is known, the objective of MS analysis is to identify and quantify modified proteoforms of the protein [19, 20]. Although hundreds of PTMs have been found on various proteins, it is common that only several *expected PTMs* are observed on the target protein. For example, expected PTMs on histone proteins include methylation, dimethylation, trimethylation, acetylation, and phosphorylation. In this study, only proteoforms with expected PTMs are considered as candidates in HomMTM spectral identification.

Let $\mathcal{Q}$ be the set of all proteoforms of an unmodified protein $P$ with expected PTMs and $\mathcal{Q}_M$ a subset of $\mathcal{Q}$ containing all the proteoforms with a molecular mass $M$. For example, when acetylation on lysine residues is the only expected PTM, the set $\mathcal{Q}$ for the protein AKGKL contains three proteoforms AK[acetylation]GKL, AKGK[acetylation]L, and AK[acetylation]GK[acetylation]L. When $M$ is the sum of the mass of the protein and the mass shift of an acetylation, $\mathcal{Q}_M$ contains only two proteoforms AK[acetylation]GKL and AKGK[acetylation]L. Let $S$ be a HomMTM spectrum with a precursor mass $M$ generated from proteoforms $Q_1, Q_2, \ldots, Q_k$ of protein $P$. The molecular masses of the $k$ proteoforms is the same to the precursor mass of $S$. That is, $Q_1, Q_2, \ldots, Q_k \in \mathcal{Q}_M$. In practice, errors in the precursor mass of $M$ are allowed, and the set $\mathcal{Q}_M$ contains all proteoforms whose molecular masses are similar to $M$ (the difference is within an error tolerance).

**Definition 2** *Given a set $\mathcal{T}$ of expected PTMs, a protein P, a HomMTM spectrum S with a precursor mass M, and a number k, the HomMTM spectral identification problem is to find k proteoforms $Q_1, Q_2, \ldots, Q_k \in \mathcal{Q}_M$ and their abundances that best explain the spectrum S.*

#### *Representing the HomMTM spectral identification problem as a graph problem*

We will formulate the HomMTM spectral identification problem as the ME$k$SF problem. The proposed method can be applied to tandem mass spectra with various fragmentation methods, such as collision-induced dissociation (CID), higher-energy collision dissociation (HCD), and electron-transfer dissociation (ETD). Here HCD tandem mass spectra are used to explain the method. Only one type of N-terminal fragment ions and one type of C-terminal fragment ions are considered in the method to simply the analysis.

Tandem mass spectra of proteoforms in top-down MS often contain high charge state fragment ions and isotopomer envelopes. The first step in interpreting these spectra is to convert a spectrum into a list of monoisotopic fragment masses using top-down spectral deconvolution tools, such as Thrash [21] and MS-Deconv [22]. In the following analysis, we assume that the spectrum $S$ is a deconvoluted tandem mass spectrum.

The target protein $P$ is represented as a sequence of amino acids $a_1a_2, \ldots, a_n$. The $i$th prefix residue mass of $P$ is the sum of the residue masses of its first $i$ amino acids, that is, $p_i = \sum_{k=1}^{i} \text{Mass}(a_k)$, where $\text{Mass}(a_k)$ is the residue mass of $a_k$. Specifically, $p_0 = 0$. The $i$th suffix residue mass of $P$ is the sum of the residue masses of its last $i$ amino acids, that is, $s_i = \sum_{k=n-i+1}^{n} \text{Mass}(a_k)$. Because of the existence of PTMs, a proteoform $Q$ in $\mathcal{Q}_M$ may have prefix residue masses different from those of $P$, and the $i$th prefix residue mass of $Q$ is the sum of $p_i$ and the mass shifts of the PTMs on the first $i$ amino acids in $Q$. Two different proteoforms in $\mathcal{Q}_M$ may have the same $i$th prefix residue mass because they have the same mass shifts on the first $i$th amino acids. For example, GK[acetylation]GKL and GKGK[acetylation]L have the same 4th prefix residue mass because the first 4 amino acids in the two proteoforms have the same PTM acetylation on different sites. Similarly, different proteoforms in $\mathcal{Q}_M$ may have the same $i$th suffix residue mass. Let $\mathcal{P}_i$ ($\mathcal{S}_i$), for $0 \le i \le n$, be the set of all different $i$th prefix (suffix) residue masses of the proteoforms in $\mathcal{Q}_M$. The $i$th prefix residue mass and the $n-i$th suffix residue mass of a proteoform in $Q$ are called *complementary masses*. Each prefix residue mass in $\mathcal{P}_i$ has a corresponding complementary suffix residue mass in $\mathcal{S}_{n-i}$. Let $\mathcal{T}_i$ be the set of expected PTMs that can occur on the $i$th amino acid in $P$. A mass $m_1$ in $\mathcal{P}_i$ is a *preceding mass* of another mass $m_2$ in $\mathcal{P}_{i+1}$ if $m_2 - m_1$ matches $\text{Mass}(a_{i+1})$ or the sum of $\text{Mass}(a_{i+1})$ and the mass shift of a PTM in $\mathcal{T}_i$.

Theoretical masses in $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_n, \mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_n$ are compared with deconvoluted fragment masses in $S$ to find matched ones. Mass shifts determined by fragment ion types are added these theoretical masses in the matching because the theoretical prefix or suffix residue masses

may have mass shifts compared with their corresponding experimental fragment masses. For example, the mass 18.015 Dalton (Da) of a water molecule is added to theoretical suffix residue masses to match experimental y-ion neutral fragment masses. The raw intensity of a prefix residue mass $m$ is the sum of intensities of neutral fragment masses in $S$ that match either $m$ or the complementary suffix residue mass of $m$, denoted by Inte($m$). The relative intensity of a prefix residue mass is the ratio between the raw intensity of the mass and the largest raw intensity of all prefix residue masses (Fig. 1a).

A directed graph $G$ containing $n+1$ layers is generated from the sets of prefix residue masses $\mathcal{P}_0, \mathcal{P}_1, \ldots, \mathcal{P}_n$ with five steps (Fig. 1b). (1) A vertex is added to the $i$th layer of $G$ for each mass in $\mathcal{P}_i$. (2) A vertex $u$ in the $i$th layer is connected to another vertex $v$ in the $i+1$th layer by a directed edge if and only if the mass corresponding to $u$ is a preceding mass of the mass corresponding to $v$. (3) The only vertex in layer 0 is labeled as the source vertex, and the only vertex in layer $n$ is labeled as the sink vertex. (4) We remove all vertices that are not on any path from the source to the sink. (5) Let $m_1, m_2, \ldots, m_k$ be the prefix masses corresponding to the remaining vertices in a layer of $G$. The capacity of the vertex corresponding to mass $m_i$ is defined as $\frac{\text{Inte}(m_i)}{\sum_{j=1}^{k} \text{Inte}(m_j)}$.

Each path from the source to the sink in $G$ corresponds to a proteoform in $\mathcal{Q}_M$, and the flow of a path corresponds to the relative abundance of the proteoform. Using this method, the HomMTM spectral identification problem

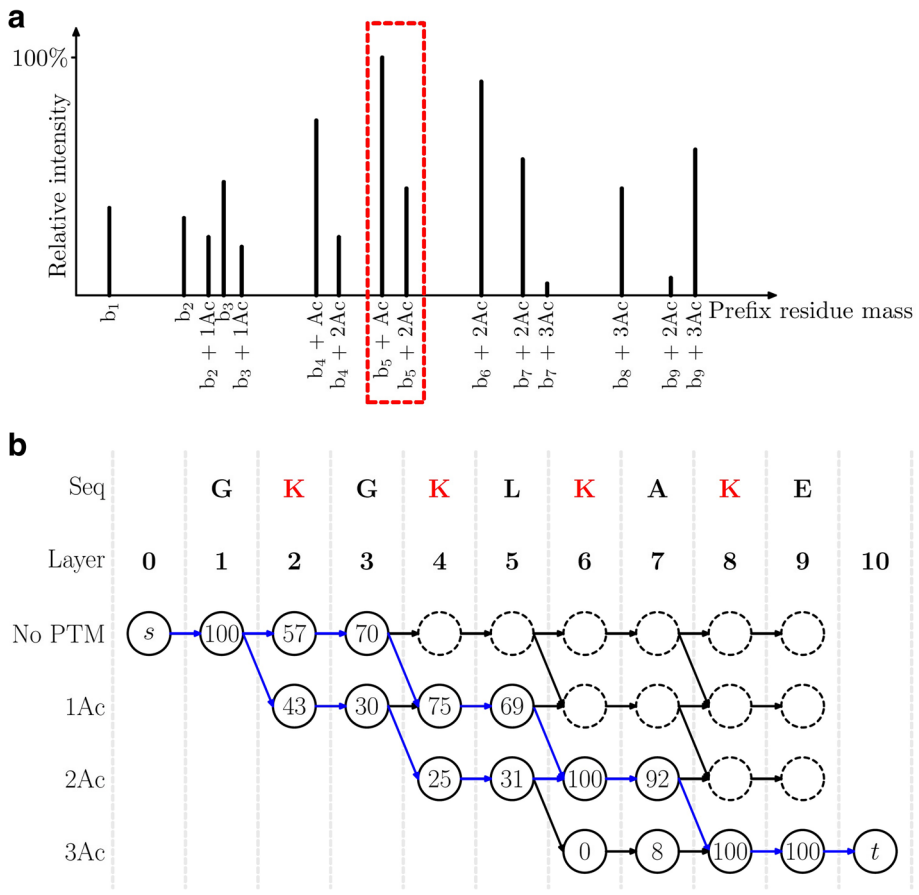


**Fig. 1** Illustration of the conversion from the HomMTM spectral identification problem to the MS*k*SF problem. A deconvoluted HomMTM spectrum generated from two modified proteoforms of the protein GKGKLKAKE with one expected PTM: acetylation on K, is used as an example. **a** Each peak corresponds to a potential prefix residue mass of a proteoform of GKGKLKAKE satisfying that the prefix residue mass or its complementary suffix residue mass matches an experimental fragment mass. Potential masses for the prefix GKGKL matched to experimental masses are shown in the red dotted box. **b** A graph with 10 layers is constructed based on the masses in $\mathcal{P}_0, \mathcal{P}_1, \ldots, \mathcal{P}_{10}$ and the peaks in (**a**). Each vertex in layer $i$, $0 \le i \le 10$, corresponds to a mass in $\mathcal{P}_i$ and those with dotted circles are removed because they are not on any path from the source to the sink. The capacity of a vertex is the ratio (shown in percentage) between the intensity of the mass and the sum of the intensities of all masses corresponding to vertices with solid circles in the same layer. The solution to the MS*k*SF problem is the two blue paths with flows 70 and 30 (in percentage), which correspond to two proteoforms GK[Acetylation]GK[Acetylation]LKAKE with relative abundance 70% and GKGK[Acetylation]LK[Acetylation]AKE with relative abundance 30%

is transformed into an ME*k*SF problem on a graph, in which the total flow value is fixed (100 was used in the experiments).

#### An algorithm for the ME2SF problem

The ME*k*SF problem is NP-hard on directed acyclic graphs when $k$ is part of the input, which can be proved by reducing from the partition problem [23]. (See Additional file 1.) Here we propose a dynamic programming algorithm for the ME*k*SF problem for $k$=2 on layered directed graphs.

A directed graph $G = (V, E)$ is a layered one if there exists a partition of its vertex set $V = \{V_1, V_2, \cdots, V_h\}$, such that $(u, v) \in E$ if and only if $u \in V_i$ and $v \in V_{i+1}$ for $1 \leq i \leq h-1$. Let $G = (\{V_1, \cdots, V_h\}, E)$ be a layered directed graph in which $V_1 = \{s\}$ and $V_h = \{t\}$. Following the terminology introduced in the studies of the M*k*SF problem, which has many applications on commodity transportation, a flow value pair $(f_1, f_2)$, $f_1, f_2 \geq 0$, is called a *packing*, and the packing is optimal for the ME2SF problem if there is an optimal ME2SF $(P_1, f_1), (P_2, f_2)$.

Koch et al. has proved that the M*k*SF problem can be solved in polynomial time on graphs with bounded treewidth, including layered directed graphs, when $k$ is a constant [16]. The method consists of two steps: the packing step finds candidates for the flow values of the $k$ paths in an optimal flow, and the routing step reports $k$ paths with the minimum error for each candidate. Similarly, in the proposed algorithm for the ME2SF problem, we first generate candidate packings that contain an optimal one, then find the best routing for each packing.

In the packing step, the total flow value $f$ is fixed, a naive approach is to enumerate all possible packings $(f_1, f_2)$ such that $f_1 + f_2 = f$. The number of candidate packings is $O(f)$, which may be an exponential function of the length of the input. Below we show that it is sufficient to consider $O(|V|)$ packings to solve the ME2SF problem.

A set $\mathcal{S}$ of candidate packings with an $O(|V|)$ size is generated as follows: (1) for each vertex $v \in V$ with $c(v) < f$, a candidate packing $(c(v), f - c(v))$ is added to $\mathcal{S}$; (2) a special packing $(f, 0)$ is added to $\mathcal{S}$. The total number of candidate packings in $\mathcal{S}$ is no large than $|V| + 1$.

We will prove the candidate set $\mathcal{S}$ contains at least one optimal packing. Let $F = (P_1, f_1), (P_2, f_2)$ be an optimal solution to the ME2SF problem, in which $V_1$ is the set of vertices in $P_1$, $V_2$ is the set of vertices in $P_2$, and $V_1 \neq V_2$. Let $v^*$ be a vertex in $(V_1 - V_2) \cup (V_2 - V_1)$ with the minimum capacity error. A vertex $v$ with $f(v) < c(v)$, $f(v) = c(v)$, $f(v) > c(v)$ is called an *under flow*, *perfect flow*, *over flow* vertex, respectively. The numbers of over flow and under flow vertices in $V_1 - V_2$ are denoted as $n_1^+$ and $n_1^-$, respectively; the numbers of over flow and under flow vertices in $V_2 - V_1$ are denoted as $n_2^+$ and $n_2^-$, respectively.

**Lemma 1** *If $v^*$ is not a perfect flow vertex, then $n_1^+ + n_2^- = n_1^- + n_2^+$. That is, the sum of the numbers of over flow vertices in $V_1 - V_2$ and under flow vertices in $V_2 - V_1$ equals the sum of the numbers of under flow vertices in $V_1 - V_2$ and over flow vertices in $V_2 - V_1$.*

*Proof* We prove the lemma by contradiction. If $n_1^+ + n_2^- < n_1^- + n_2^+$, then we increase the flow value of $P_1$ by $\delta = \varepsilon(v^*)$ and decrease the flow value of $P_2$ by $\delta$ to obtain a new flow $(P_1, f_1 + \delta), (P_2, f_2 - \delta)$. By increasing the flow value in $P_1$, the error of each over flow vertex in $V_1 - V_2$ increases by $\delta$, and the error of each under flow vertex in $V_1 - V_2$ decreases by $\delta$ because $\delta = \varepsilon(v^*)$ is the smallest error of the vertices in $(V_1 - V_2) \cup (V_2 - V_1)$. By decreasing the flow value in $P_2$, the error of each over flow vertex in $V_2 - V_1$ decreases by $\delta$, and the error of each under flow vertex in $V_2 - V_1$ increases by $\delta$. In addition, the errors of the vertices not in $(V_1 - V_2) \cup (V_2 - V_1)$ do not change. As a result, the error of the new flow is $\varepsilon(F) + \left(n_1^+ + n_2^- - n_1^- - n_2^+\right)\delta$, which is smaller than the error of $F$. This is a contradiction. Similarly, if $n_1^+ + n_2^- > n_1^- + n_2^+$, then the error of the flow $(P_1, f_1 - \delta), (P_2, f_2 + \delta)$ is smaller than the error of $F$, which is a contradiction. □

**Theorem 1** *The candidate set $S$ contains at least one optimal packing of $G$.*

*Proof* Let $F = (P_1, f_1), (P_2, f_2)$ be an optimal solution to the ME2SF problem. We consider two cases: (1) $P_1$ and $P_2$ are the same and (2) $P_1$ and $P_2$ are different. In the first case, the two paths are the same, then $(P_1, f), (P_2, 0)$ is an optimal solution and $(f, 0) \in \mathcal{S}$ is an optimal packing. In the second case, we will prove that there exists an optimal solution $F' = \left(P_1, f_1'\right), \left(P_2, f_2'\right)$ such that $\left(f_1', f_2'\right)$ or $\left(f_2', f_1'\right)$ is in $\mathcal{S}$.

Suppose $P_1$ and $P_2$ are different and $(V_1 - V_2) \cup (V_2 - V_1)$ is not empty. Let $v^*$ be a vertex with the minimum error in $(V_1 - V_2) \cup (V_2 - V_1)$. Without loss of generality, we assume that $v^* \in V_1 - V_2$. If $v^*$ is a perfect flow edge, then $F' = F$ and $(f_1, f_2) = (c(v^*), f - c(v^*)) \in S$. Otherwise, based on Lemma 1, $n_1^+ + n_2^- = n_1^- + n_2^+$. By changing the flow values $(f_1, f_2)$ to $(c(v^*), f - c(v^*))$, we obtain a new flow $F' = (P_1, c(v^*)), (P_2, f - c(v^*))$. The difference between the errors of $F$ and $F'$ is $\left(n_1^+ + n_2^- - n_1^- - n_2^+\right)\varepsilon = 0$. As a result, $F'$ is an optimal solution, and $(c(v^*), f - c(v^*)) \in \mathcal{S}$ is an optimal packing. □

In the routing step, we propose a dynamic programming algorithm for finding a flow $(P_1, f_1, ), (P_2, f_2)$ for a packing $(f_1, f_2)$ such that the error of the flow is minimized. We first introduce partial flows that are used in the routing algorithm. A path pair with flows $(P_1, f_1), (P_2, f_2)$ is called a partial 2-splittable $s$-$t$-flow if $P_1$ and $P_2$ start at the source $s$. The two paths in the partial flow may not end at the

sink $t$. The error of a partial flow is defined similarly as the error of a 2-splittable $s$-$t$-flow.

For each ordered vertex pair $(v_1, v_2)$ ($v_1$ and $v_2$ may be the same) in a layered directed graph $G = (\{V_1, V_2, \ldots, V_h\}, E)$, we define $D(v_1, v_2)$ as the minimum error of all partial 2-splittable flows $(P_1, f_1), (P_2, f_2)$ such that $P_1$ ends at $v_1$ and $P_2$ ends at $v_2$. A vertex pair $\left(v_1', v_2'\right)$ ($v_1'$ and $v_2'$ may be the same) *precedes* vertex pair $(v_1, v_2)$ if $\left(v_1', v_1\right), \left(v_2', v_2\right) \in E$. The error of an ordered vertex pair $(v_1, v_2)$ for a packing $(f_1, f_2)$ is defined as

$$\varepsilon(v_1, v_2) = \begin{cases} |c(v_1) - f_1 - f_2| & \text{if } v_1 = v_2; \\ |c(v_1) - f_1| + |c(v_2) - f_2| & \text{otherwise.} \end{cases}$$

Let $T(v_1, v_2)$ be the set of all precedent pairs of $(v_1, v_2)$. We use a dynamic programming algorithm to fill out $D(v_1, v_2)$ for all vertex pairs $v_1, v_2$ in the same layer. The recurrence function for computing $D(v_1, v_2)$ is

$$D(v_1, v_2) = \min_{(v_1', v_2') \in T(v_1, v_2)} D(v_1', v_2') + \varepsilon(v_1, v_2).$$

After obtaining the value $D(t, t)$ for the sink, we use backtracking to find the best path pair for the ME2SF problem. The time complexity of the routing algorithm is $O(l^4 h)$ where $l$ is the largest number of vertices in a layer and $h$ is the number of layers in $G$. Since $O(|V|)$ packings are searched for finding the best solution, the time complexity of the algorithm for the ME2SF problem is $O(l^4 h|V|)$. In practice, the value $l$ is not large in most cases, and the proposed algorithm is efficient for the ME2SF problem.

## Results

We implemented the dynamic programming algorithm for the ME2SF problem in C++ and tested it on a top-down MS/MS data set of the human histone H4 protein. The experiments were performed on a Linux server with Intel(R) Xeon(R) E5-2680 2.5 GHz CPU.

### Data set

Core histone proteins collected from normal human dermal fibroblasts were separated using a 2-dimensional reverse phase hydrophilic interaction liquid chromatography (RP-HILIC) system. Histone H4 isolated in the first dimension of the separation was analyzed using an LTQ Orbitrap Velos with a resolution of 60k for MS and MS/MS spectra. In total, 1,626 CID and 1,626 ETD spectra were acquired. Details of the experiment can be found in ref. [9].

### Proteoform identification

All top-down tandem mass spectra were deconvoluted using MS-Deconv [22]. In the proposed algorithm, the error tolerances for precursor and fragment masses were set to 15 parts-per million (ppm); the maximum mass difference between the molecular mass of the unmodified protein sequence and the precursor mass of the spectrum was set to 200 Da; five PTMs were treated as expected ones (Table 1). Spectral deconvolution often introduces $\pm 1$ Da errors into precursor masses of top-down tandem mass spectra. To address this problem, $\pm 1$ Da errors were also allowed in matching precursor masses to the molecular masses of proteoforms.

With a cutoff of 10 matched fragment ions, the proposed algorithm identified 441 spectra matched to single proteoforms and 184 spectra matched to proteoform pairs. The running time was about 25 minutes and the memory requirement was about 32 GB. If the proteoform pair matched to a spectrum provides explanation for many fragment ions that are not explained by one proteoform, it is highly possible that the spectrum is a HomTMT spectrum. For 39 of the 184 spectra, the proteoforms pairs have at least 10 more explained fragment ions than the single high abundance proteoforms in these pairs. In addition, for 26 of the 184 spectra, the proteoform pairs have at least 20% more explained peak intensity than the single high abundance proteoforms.

### Comparison with MS-Align-E

MS-Align-E [9] was employed to align the histone H4 protein with the deconvoluted tandem mass spectra in the data set. With the same error tolerances and expected PTMs described in the previous subsection, MS-Align-E identified 1037 proteoform-spectrum-matches with at least 10 matched fragment ions. The main reason that MS-Align-E identified more spectra is that unexpected PTMs are allowed in MS-Align-E, but not in the proposed method. The 184 spectra matched to proteoform pairs by the proposed method were all identified by MS-Align-E. For these spectra, we compared the single proteoforms reported by MS-Align-E and the proteoform pairs reported by the proposed method in the number of matched fragment ions and the explained peak intensities. Compared with MS-Align-E, the proposed method increased the number of matched fragment ions by at least 10 for 43 spectra (Fig. 2) and increased the explained peak intensities by at least 20% for 26 spectra, demonstrating

**Table 1** Five expected PTMs are allowed in the identification and quantification of histone H4 proteoforms

| PTM | Monoisotopic mass (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 |

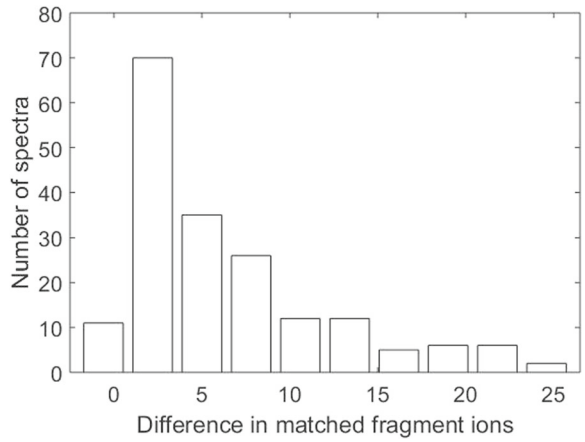


**Fig. 2** Comparison of the numbers of matched fragment ions. The numbers of matched fragment ions are compared for the 184 spectra identified by both the proposed method and MS-Align-E. For each spectrum, the difference between the number of fragment ions matched to the proteoform pair reported by the proposed method and that matched to the single proteoform reported by MS-Align-E is computed

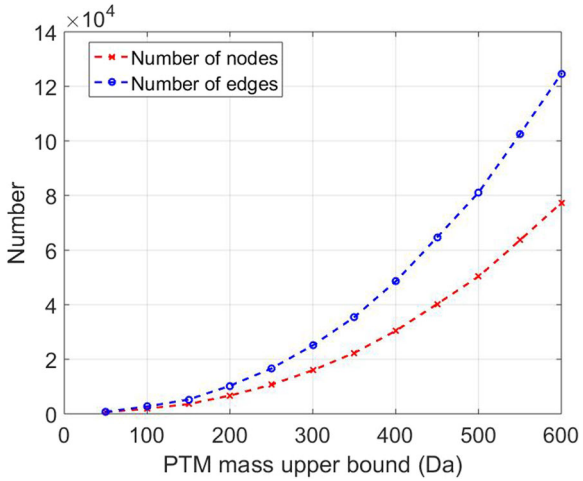


**Fig. 3** The sizes of graphs used in HomMTM spectral interpretation. The numbers of vertices and edges in the graph generated from the histone H4 protein and five PTMs (acetylation, methylation, dimethylation, trimethylation, phosphorylation) increase significantly when the bound for the sum of mass shifts introduced by PTMs increases from 50 Da to 600 Da

that these proteoform pairs better explain the spectra than the single proteoforms.

### Parameter selection

The size of the graph generated from a protein sequence and a set of expected PTMs may be huge due to the combination of PTMs. We can reduce the size of the graph by introducing a bound for the sum of mass shifts introduced by PTMs. When the bound increases from 50 Da to 600 Da, the size of the graph generated from the histone H4 protein and the five expected PTMs increases significantly: the number of vertices increases from 606 to 77,246; the number of edges increases from 761 to 124,633 (Fig. 3). The size of the graph is proportional to $hq^t$, where $h$ is the number of layers in the graph, $q$ is the largest number of PTM sites in a proteoform, and $t$ is the number of expected PTM types. The bound for the sum of mass shifts of PTMs is used to limit the number $q$. In practice, when the number of expected PTM types $t$ is 1 or 2, the size of the graph increases slowly with respect to $q$ and a large bound 1000 or 1500 Da can be used to allow more PTM sites in a proteoform. When the number $t$ is 4 or 5, the size of the graph increases rapidly with respect to $q$ and a small bound 500 or 600 Da should be used to guarantee the efficiency of the algorithm.

## Discussion

Because experimental mass spectra contain many noise peaks and miss many fragment peaks, it is not easy to confidently identify more than 2 proteoforms from a HomMTM spectrum. Therefore, we focus on the HomMTM spectral identification problem for $k = 2$. The proposed algorithm in the routing step can be extended to the case with $k > 2$, but the one in the packing step cannot. A trivial algorithm for the packing step is to enumerate all combinations of flow values for the $k$ paths, and the number of candidate packings is $O(f^{k-1})$, where $f$ is the total flow value. Coupled with the dynamic programming algorithm for the routing step, the time complexity of the combined method is $O\left(l^{2k}hf^{k-1}\right)$.

## Conclusions

We formulated the HomMTM spectral identification problem as the ME$k$SF problem on graphs and proposed an efficient algorithm for solving the ME2SF problem on layered directed graphs. The experiments on the histone H4 data set demonstrated that the proposed algorithm is capable of identifying many top-down MTM spectra and gives better explanation for these spectra using proteoform pairs.

## Additional file

**Additional file 1:** Supplementary material. (PDF 162 kb)

### Abbreviations

CID: Collision-induced dissociation; Da: Dalton; ETD: Electron-transfer dissociation; GB: Gigabyte; HCD: Higher-energy collision dissociation; HetMTM spectrum: Heterogeneous multiplexed tandem mass spectrum; HomMTM spectrum: Homogeneous multiplexed tandem mass spectrum; LTQ: Linear trap quadrupole; ME2SF: Minimum error 2-splittable flow; ME$k$SF: Minimum error $k$-splittable flow; M$k$SF: Maximum $k$-splittable flow; MS: Mass spectrometry; MS/MS: Tandem mass spectrometry; MTM spectrum: Multiplexed tandem mass spectrum; NP-hard: Non-deterministic polynomial-time hard; PTM: Post-translational modification; ppm: Parts-per million; RP-HILIC: Reverse phase hydrophilic interaction liquid chromatography

**Funding**
This work and publication of this article were supported by the National Institute of General Medical Sciences, National Institutes of Health (NIH) through Grant R01GM118470.

**Availability of data and materials**
All data used in this paper are available at Massive (http://massive.ucsd.edu) with the identification number MSV000082054.

**About this supplement**
This article has been published as part of *BMC Bioinformatics* Volume 19 Supplement 9, 2018: Selected articles from the 13th International Symposium on Bioinformatics Research and Applications (ISBRA 2017): bioinformatics. The full contents of the supplement are available online at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-19-supplement-9.

**Authors' contributions**
KZ and XL formulated the HomMTM spectral identification problem and designed the graph-based algorithm. KZ implemented the algorithm and carried out the experiments. All authors read and approved the final manuscript.

**Ethics approval and consent to participate**
Not applicable.

**Consent for publication**
Not applicable.

**Competing interests**
The authors declare that they have no competing interests.

## Publisher's Note



**Author details**
[1]Department of Computer Science, Indiana University Bloomington, 700 N. Woodlawn Avenue, 47408 Bloomington, IN, USA. [2]Department of BioHealth Informatics, Indiana University-Purdue University Indianapolis, 719 Indiana Avenue, 46202 Indianapolis, IN, USA. [3]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, 410 W. 10th Street, 46202 Indianapolis, IN, USA.

Published: 13 August 2018

**References**

1. DiMaggio Jr PA, 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. Mol Cell Proteomics. 2009;8:2527–43.
2. Wang J, Perez-Santiago J, Katz JE, Mallick P, Bandeira N. Peptide identification from mixture tandem mass spectra. Mol Cell Proteomics. 2010;9:1476–85.
3. Wang J, Bourne PE, Bandeira N. MixGF: spectral probabilities for mixture spectra from more than one peptide. Mol Cell Proteomics. 2014;13:3688–97.
4. Cosgrove MS, Wolberger C. How does the histone code work?. Biochem Cell Biol. 2005;83:468–76.
5. Strahl BD, Allis CD. The language of covalent histone modifications. Nature. 2000;403:41–5.
6. Distler U, Kuharev J, Navarro P, Levin Y, Schild H, Tenzer S. Drift time-specific collision energies enable deep-coverage data-independent acquisition proteomics. Nat Methods. 2014;11:167–70.
7. Rost HL, Rosenberger G, Navarro P, Gillet L, Miladinovic SM, Schubert OT, Wolski W, Collins BC, Malmstrom J, Malmstrom L, Aebersold R. OpenSWATH enables automated, targeted analysis of data-independent acquisition ms data. Nat Biotechnol. 2014;32:219–23.
8. 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. Mol Cell Proteomics. 2012;11:111–008524.
9. Liu X, Hengel S, Wu S, Tolić N, Paša-Tolić L, Pevzner PA. Identification of ultramodified proteins using top-down tandem mass spectra. J Proteome Res. 2013;12:5830–8.
10. Sun RX, Luo L, Wu L, Wang RM, Zeng WF, Chi H, Liu C, He SM. pTop 1.0: A high-accuracy and high-efficiency search engine for intact protein identification. Anal Chem. 2016;88:3082–90.
11. Kou Q, Xun L, Liu X. TopPIC: a software tool for top-down mass spectrometry-based proteoform identification and characterization. Bioinformatics. 2016;32:3495–7.
12. 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. 2017;33:1309–16.
13. Baier G, Köhler E, Skutella M. On the $k$-splittable flow problem. In: Algorithms ESA 2002. Lecture Notes in Computer Science. vol. 2461. Berlin Heidelberg: Springer; 2002. p. 101–13.
14. Baier G, Köhler E, Skutella M. The $k$-splittable flow problem. Algorithmica. 2005;42:231–48.
15. Koch R, Spenke I. Complexity and approximability of $k$-splittable flows. Theor Comput Sci. 2006;369:338–47.
16. Koch R, Skutella M, Spenke I. Maximum $k$-splittable $s, t$-flows. Theor Comput Syst. 2008;43:56–66.
17. Caramia M, Sgalambro A. An exact approach for the maximum concurrent $k$-splittable flow problem. Optim Lett. 2008;2:251–65.
18. Caramia M, Sgalambro A. A fast heuristic algorithm for the maximum concurrent $k$-splittable flow problem. Optim Lett. 2010;4:37–55.
19. Moradian A, Kalli A, Sweredoski MJ, Hess S. The top-down, middle-down, and bottom-up mass spectrometry approaches for characterization of histone variants and their post-translational modifications. Proteomics. 2014;14:489–97.
20. Yuan ZF, Arnaudo AM, Garcia BA. Mass spectrometric analysis of histone proteoforms. Annu Rev Anal Chem. 2014;7:113–28.
21. Horn DM, Zubarev RA, McLafferty FW. Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. J Am Soc Mass Spectrom. 2000;11:320–32.
22. 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. Mol Cell Proteomics. 2010;9:2772–82.
23. Korf RE. A complete anytime algorithm for number partitioning. Artif Intell. 1998;106:181–203.