Bioinformatics, 2016, 1–7
doi: 10.1093/bioinformatics/btw307
Advance Access Publication Date: 14 May 2016
Original Paper

OXFORD

Sequence analysis

Top-down analysis of protein samples by *de novo* sequencing techniques

Kira Vyatkina[1,2,*], Si Wu[3], Lennard J. M. Dekker[4], Martijn M. VanDuijn[4], Xiaowen Liu[5,6], Nikola Tolić[7], Theo M. Luider[4], Ljiljana Paša-Tolić[7] and Pavel A. Pevzner[2,8,*]

[1]Algorithmic Biology Laboratory, Saint Petersburg Academic University, St Petersburg, Russia, [2]Center for Algorithmic Biotechnology, Institute of Translational Biomedicine, Saint Petersburg State University, St Petersburg, Russia, [3]Department of Chemistry and Biochemistry, University of Oklahoma, Norman, OK, USA, [4]Department of Neurology, Erasmus University Medical Center, Rotterdam, The Netherlands, [5]Department of BioHealth Informatics, Indiana University-Purdue University Indianapolis, Indianapolis, IN, USA, [6]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN, USA, [7]Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, WA, USA and [8]Department of Computer Science and Engineering, University of California, San Diego, CA, USA

*To whom correspondence should be addressed.

Associate Editor: Burkhard Rost

Received on November 9, 2015; revised on March 31, 2016; accepted on May 9, 2016

Abstract

Motivation: Recent technological advances have made high-resolution mass spectrometers affordable to many laboratories, thus boosting rapid development of top-down mass spectrometry, and implying a need in efficient methods for analyzing this kind of data.

Results: We describe a method for analysis of protein samples from top-down tandem mass spectrometry data, which capitalizes on *de novo* sequencing of fragments of the proteins present in the sample. Our algorithm takes as input a set of *de novo* amino acid strings derived from the given mass spectra using the recently proposed Twister approach, and combines them into *aggregated strings* endowed with offsets. The former typically constitute accurate sequence fragments of sufficiently well-represented proteins from the sample being analyzed, while the latter indicate their location in the protein sequence, and also bear information on post-translational modifications and fragmentation patterns.

Availability and Implementation: Freely available on the web at http://bioinf.spbau.ru/en/twister.

Contact: vyatkina@spbau.ru or ppevzner@ucsd.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Top-down mass spectrometry is a recently emerged technology for analyzing intact proteins, which is particularly suitable for detecting post-translational modifications (PTMs). Among its most important applications is characterization of therapeutic antibodies (Bondarenko *et al.*, 2009; Fornelli *et al.*, 2012; Kellie *et al.*, 2010), as part of the process of drug development. Nowadays, modern instruments allow to rapidly acquire vast amount of high-quality top-down data, but efficient algorithms for its processing are still in high demand.

One of the key tasks of mass spectrometry-based proteomics is *de novo* sequencing of peptides and proteins from tandem (MS/MS) mass spectrometry data. In the last two decades, several algorithms have been proposed for *de novo* peptide sequencing from bottom-up data, and a few powerful tools appeared, including PEAKS (Ma *et al.*, 2003), PepNovo (Frank and Pevzner, 2005), pNovo

(Chi *et al.*, 2010), Lutefisk (Taylor and Johnson, 1997), Sherenga (Dancik *et al.*, 1999), Novor (Ma, 2015) and the program UVnovo (Robotham *et al.*, 2016) recently introduced for the case of 351 nm ultraviolet photodissociation (UVPD) mass spectra. Other achievements comprise methods for complete protein sequencing using multiple enzyme digest and assembly by sequence overlap, possibly referring to a homologous protein sequence (Bandeira *et al.*, 2004, 2007, 2008; Castellana *et al.*, 2010; Liu *et al.*, 2009), and a number of alternative approaches benefit from complementarity of mass spectra acquired using different fragmentation techniques from either peptides (Bertsch *et al.*, 2009; Chi *et al.*, 2013; Datta and Bern, 2008; Guthals *et al.*, 2013; He and Ma, 2010; Savitski *et al.*, 2005) or intact proteins (Horn *et al.*, 2000).

Yet to the best of our knowledge, only three published algorithms make any use of top-down data to achieve the above-mentioned goal, and namely, the one from (Horn *et al.*, 2000), which, however, did not result in a publicly available tool, TBNovo (Liu *et al.*, 2014) that processes combined MS/MS datasets and utilizes top-down spectra as a scaffold for assembling bottom-up spectra, and our recently developed approach called Twister (Vyatkina *et al.*, 2015). At the same time, primarily due to increasing accessibility of hybrid Orbitrap Fourier transform mass spectrometers, the top-down technology is nowadays rapidly gaining popularity, thus enforcing the need in efficient approaches to *de novo* peptide and protein sequencing from top-down spectra *alone*.

This work represents a continuation of our research summarized in (Vyatkina *et al.*, 2015), where we introduced a method that extracts from deconvoluted MS/MS spectra a number of accurate sequence fragments of the proteins contained in a sample. Here we describe an algorithm that takes as input the fragments obtained this way, and combines them to produce a set of so-called *aggregated strings*, each endowed with direct and reversed offsets intended to match the masses of the two terminal fragments preceding and following it in the underlying protein sequence. Thereby we obtain longer sequence fragments of sufficiently well-represented proteins, among which should be the target ones (but can be also abundant enough contaminants). The number of aggregated strings is substantially smaller than that of the original fragments, thus allowing to see at a glance distinct components of the sample. Moreover, their associated offsets can provide extra insight into what was the composition of the sample at the level of proteoforms, which post-translational modifications (PTMs) occurred in the sequence, or which fragment ions were measured in the experiment. This immediately turns the suggested approach into a handy stand-alone method for analyzing protein samples.

In what follows, we outline the procedure for computing aggregated strings, illustrate the potential of the method by experimental results obtained for top-down MS/MS datasets for the standard protein carbonic anhydrase 2 (CAH2) and the Fab region of alemtuzumab, and indicate directions for further research.

The proposed approach is implemented in a software tool Twister available at http://bioinf.spbau.ru/en/twister, which also incorporates the method introduced in (Vyatkina *et al.*, 2015), and thus, takes as input a set of deconvoluted MS/MS spectra.

2 Methods and algorithms

2.1 Datasets

We used the top-down datasets for CAH2 and the light chain of alemtuzumab published in (Vyatkina *et al.*, 2015) and available at http://bioinf.spbau.ru/en/twister. In brief, intact CAH2, and alemtuzumab digested with papain and reduced, were analyzed by a reversed-phase liquid chromatography (RPLC) system coupled online with a Thermo LTQ Orbitrap Elite and Velos, respectively. MS and MS/MS spectra were collected at a resolution of 240k and 120k, respectively, in case of CAH2, and 100k and 60k, respectively, in case of alemtuzumab. The CAH2 and alemtuzumab datasets comprised 3, 033 ETD, 3, 363 CID and 3, 437 HCD top-down MS/MS spectra, and 4, 962 ETD and 4, 931 HCD top-down MS/MS spectra, respectively.

2.2 Generation of aggregated strings

When describing our approach, we will refer to the notion of a *peptide sequence tag* (Mann and Wilm, 1994), or simply *tag*, being a short amino acid sequence with an associated offset. A tag of length k, or k-tag, is defined by $k+1$ peaks $p_1, \ldots, p_{k+1}$ from a spectrum S, such that each two consecutive ones are separated by the mass of an amino acid. Thus, a k-tag t has an amino acid sequence $s(t) = a_1 \ldots a_k$ and an offset $o(t)$ equal to the mass $\mathrm{Mass}(p_1)$ of the leftmost peak p_1. The score of a k-tag t is defined as $\mathrm{Score}(t) = \sum_{i=1}^{k+1} I(p_i)$, where $I(p_i)$ denotes the intensity of a peak p_i, for $1 \leq i \leq k+1$. For an amino acid a_i of $s(t)$, we define its score as $\mathrm{Score}_t(a_i) = I(p_i) + I(p_{i+1})$, where $1 \leq i \leq k$. In what follows, the tag length k is assumed to be fixed.

Two k-tags t_1 and t_2 originating from two distinct spectra, and having the same amino acids sequence and approximately (up to a predefined tolerance) the same offset can be merged together. Without loss of generality, assume that $\mathrm{Score}(t_1) > \mathrm{Score}(t_2)$; then for the ith underlying peak p_i^* of the resulting tag t^*, we have $\mathrm{Mass}(p_i^*) = \mathrm{Mass}(p_i^1)$ and $I(p_i^*) = I(p_i^1) + I(p_i^2)$, where p_i^1 and p_i^2 denotes the ith defining peak of t_1 and t_2, respectively, and $1 \leq i \leq k+1$. This gluing procedure extends to the case of three or more such k-tags in a straightforward way.

The input for our method constitutes a list $\mathcal{N}$ of amino acid strings endowed with offsets, derived from the deisotoped and charge state deconvoluted, often referred to simply as *deconvoluted*, MS/MS spectra as explained in (Vyatkina *et al.*, 2015). In brief, the input spectra are first preprocessed, which comprises merging nearby peaks (i.e. the ones with approximately the same masses deconvoluted from different charge states), optional peak reflection followed by merging nearby peaks (i.e. the ones due to pairs of complementary fragment ions in the original spectra), and optional water-loss ion elimination. Subsequently, the *de novo* strings are assembled from high-quality k-tags extracted from the resulting spectra, consistent with each other in terms of both amino acid sequences and offsets (Figs. 1a and 1b). This task is accomplished through constructing for the obtained set of k-tags a T-Bruijn graph (Vyatkina *et al.*, 2015), which is a modification for the case of tags of an A-Bruijn graph (Pevzner *et al.*, 2004) widely used in genomics, and extracting an optimal path from each its connected component. Thus, the k-tags together giving rise to a single *de novo* string often come from several distinct spectra. Formally, for a string $s = a_1 \ldots a_m$ from $\mathcal{N}$, the set $\mathcal{T}(s)$ of its underlying k-tags represents a disjoint union of non-empty sets $\mathcal{T}_1(s), \mathcal{T}_2(s), \ldots, \mathcal{T}_{m-k+1}(s)$ of k-tags, such that

- all the tags composing the set $\mathcal{T}_i(s)$ have the same amino acid sequence $a_i a_{i+1} \ldots a_{i+k-1}$ and approximately the same offset $o_i(s)$, and all such tags derived from the input spectra are contained in $\mathcal{T}_i(s)$, where $1 \leq i \leq m-k+1$, and
- $o_j(s) = o_{j-1}(s) + \mathrm{Mass}(a_{j-1})$, where $1 < j \leq m-k+1$.

The string s is assigned an offset $o(s) = o_1(s)$. The multiplicity of $o(s)$ is defined as $\mu(o(s)) = \mu(s) = |\mathcal{T}(s)|$.

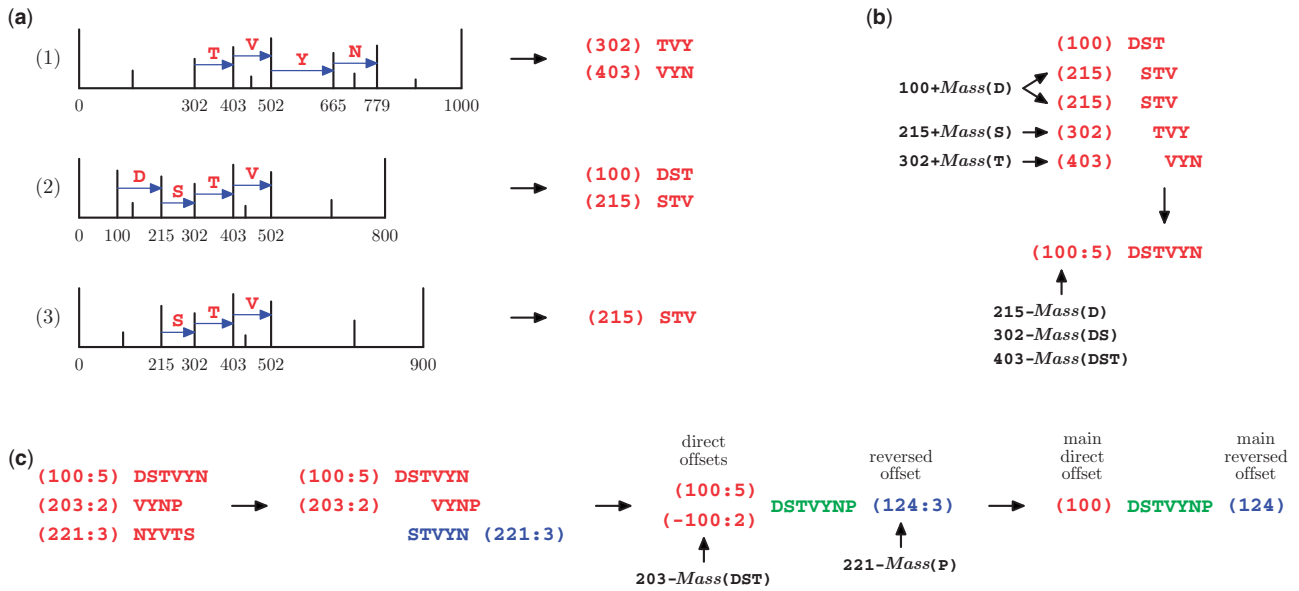

Fig. 1. Overview of the entire method for aggregated strings generation. (**a**) Extraction of 3-tags from the input set of deconvoluted MS/MS spectra. (**b**) Assembly of a *de novo* string from five 3-tags consistent with each other in terms of both amino acid sequences and offsets. The resulting *de novo* string 'DSTVYN' is assigned the mass offset of 100 with multiplicity 5. (**c**) Formation of an aggregated string from three *de novo* string endowed with offsets. The resulting aggregated string 'DSTVYNP' has the main direct and reversed offset of 100 and 124, respectively, which leads to the precursor mass estimate $PM_{\text{DSTVYNP}} = 100 + \text{Mass(DSTVYNP)} + 124 = 1000$

An interpretation of the procedure of generating s from $\mathcal{T}(s)$ consistent with the method from (Vyatkina *et al.*, 2015) is the following: first, the tags from each set $\mathcal{T}_i(s)$ are glued together, thus giving rise to a tag t_i^*, where $1 \le i \le m-k+1$, and next, the *de novo* string s is traced out. Consequently, we define the score of an amino acid a_i of s as the sum of its scores in the tags resulting from the gluing that cover it, normalized by the number of those: $\text{Score}_s(a_i) = \left(\sum_{t^* \in \mathcal{T}^*(a_i)} \text{Score}_{t^*}(a_i)\right)/|\mathcal{T}^*(a_i)|$, where $\mathcal{T}^*(a_i) = \{t_j^* | a_i \in s(t_j^*), 1 \le j \le m-k+1\}$.

For example, the score of the amino acid 'S' of the *de novo* string 'DSTVYN' shown in Figure 1b would be computed from the 3-tags labeled with 'DST' and 'STV', and the 3-tag labeled with 'STV' derived from spectrum (2) and (3) on Figure 1a, respectively, as $\{[(I^2_{215}+I^2_{302})]+[(I^2_{215}+I^2_{302})+(I^3_{215}+I^3_{302})]\}/2$, where I^j_M denotes the intensity of the peak with mass M from the jth spectrum, and the first and second term in square brackets corresponds to the tag string 'DST' and 'STV', respectively.

The score of s is defined as $\text{Score}(s) = \sum_{i=1}^{m} \text{Score}_s(a_i)$. Note that this scoring is different from the one used in (Vyatkina *et al.*, 2015) to order the output *de novo* strings, which equals $\sum_{t \in \mathcal{T}(s)} \text{Score}(t)$. While such tag-based scoring is appropriate for working with the *de novo* strings, it cannot be applied to the aggregated strings we are going to construct, in particular, since during their formation, we can encounter amino acid sequences not entirely covered with the k-mers labeling the tags generated from the input spectra. Therefore, for the aggregated strings, a scoring needs to be defined on the basis of individual amino acids rather than tags, and for the purpose of unification, at this stage we use an analogous scoring for the *de novo* strings as well. (The score of an aggregated string is formally introduced in Appendix.)

Observe that the sequence $s(t)$ of a correct k-tag t may correspond to a *reversed* k-mer of the underlying protein sequence (this will be the case e.g. for a correct tag defined by y-ions from a CID or HCD spectrum), and groups of such tags may produce reversed copies of longer fragments of the protein sequence. As a consequence, along with the amino acid strings from $\mathcal{N}$, we will have to consider their reflected counterparts. For an amino acid sequence z, its reversed copy will be further denoted by $\overline{z}$.

Construction of the aggregated strings proceeds in two steps: at the initialization stage, the amino acid sequences on their own are iteratively formed, and subsequently, their associated offsets are computed. An overview of the entire method and the initialization stage is provided in Figures 1 and 2, respectively. Note that the toy example shown in Figure 1c is intended solely to give a general idea of how an aggregated string is formed, and therefore, many details mentioned in the next section and Figure 2 are omitted in it.

2.2.1 Initializing aggregated strings

Throughout the process, we maintain a set $\mathcal{I}$ of *active strings*. Initially, $\mathcal{I} = \mathcal{N}$; at any further moment, $\mathcal{I}$ consists of all the aggregated strings obtained so far and all the strings from $\mathcal{N}$ that have not contributed to any of those. At each step, a number of strings from $\mathcal{I}$ are selected to produce one or a few new aggregated strings, and then the former get replaced in $\mathcal{I}$ by the latter.

In addition, we maintain the set $\mathcal{K}_\mathcal{I}$ of all k-mers of the strings from $\mathcal{I}$. The score of a k-mer $\rho \in \mathcal{K}_\mathcal{I}$ is defined as the sum of values contributed by the strings s from $\mathcal{I}$ that contain ρ as a substring: s contributes 1 if it represents an aggregated string, and the number n_s of the *de novo* strings from $\mathcal{N}$ with the amino acid sequence s otherwise. At each iteration, the top-scoring k-mer ρ_0 from $\mathcal{K}_\mathcal{I}$ is selected. Subsequently, the Algorithm (1) forms the set $\mathcal{J} \subseteq \mathcal{I}$ of all the active strings containing ρ_0 or $\overline{\rho_0}$, (2) builds the best alignment of those, thereby reversing the strings containing $\overline{\rho_0}$ but not ρ_0, and possibly reversing those containing both ρ_0 and $\overline{\rho_0}$, (3) if needed, splits the aligned strings into two or more clusters, such that within each cluster, any two strings match each other well and (4) from each cluster, computes the consensus string with respect to the amino acid scores, and possibly eliminates its unreliable terminal amino acids, and corrects the inner amino acids of the

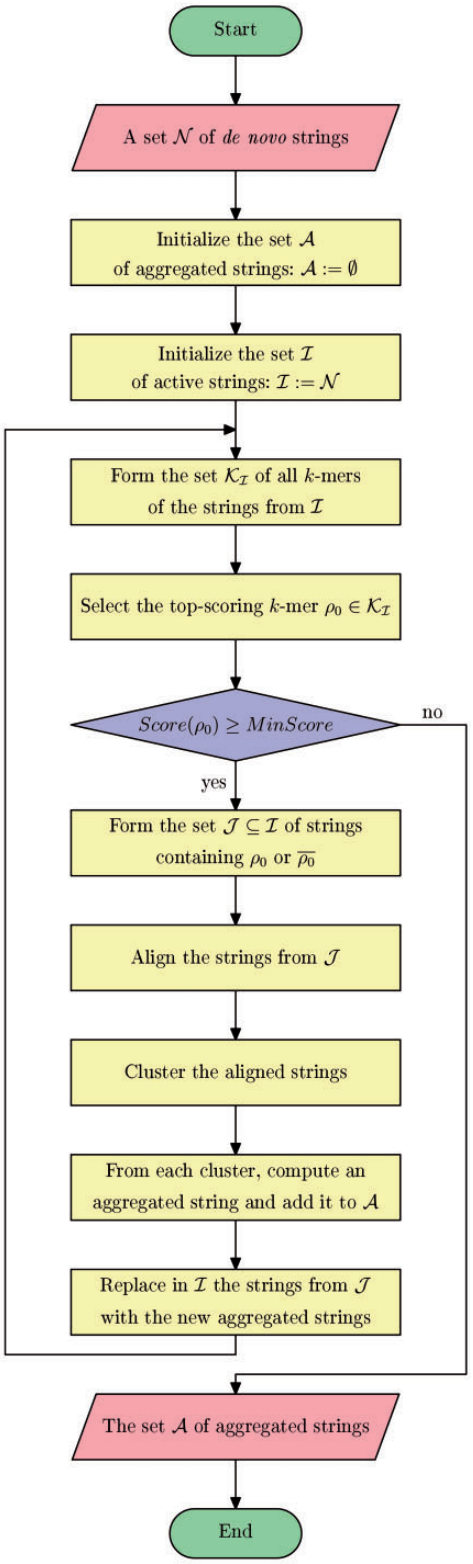

Fig. 2. Overview of the procedure generating the set of aggregated strings from a set of *de novo* amino acid strings obtained using the Twister approach (Vyatkina *et al.*, 2015)

remaining string. As a result, a new aggregated string is obtained from each cluster. These stages are illustrated in Figure 3, and detailed description of the respective procedures is provided in Supplementary Materials.

In case of ties at time of selecting an object (e.g. if two or more k-mers have the highest score, while the top-scoring one is needed), the first encountered is picked up.

The process terminates when the score of ρ_0 drops below a certain threshold *MinScore*.

This general scheme is summarized in Supplementary Algorithm 1.

2.2.2 Assigning offsets

When the final set of aggregated strings has been obtained, each of the latter gets endowed with a set of offsets (see Fig. 1c).

Consider an aggregated string **a**. Slightly abusing the notation, we will also refer by **a** to its amino acid sequence. Let $\mathcal{J}_\mathbf{a} \subseteq \mathcal{N}$ denote the set of strings, the alignment of which produced **a**. Let us partition $\mathcal{J}_\mathbf{a}$ into two subsets $\mathcal{J}_\mathbf{a}^{dir}$ and $\mathcal{J}_\mathbf{a}^{rev}$ of strings that became part of the aligned string set in their original (direct) and reversed form, respectively. Each string $s \in \mathcal{J}_\mathbf{a}^{dir}$ contributes to **a** a direct offset o_s^{dir} obtained as follows. Let $\mathbf{a} = a_1 \dots a_n$, and let $s = b_1 \dots b_m$. Suppose that $b_i \dots b_{i+k-1}$ is the first k-mer of s that matches **a** starting from position j, i.e. $b_i \dots b_{i+k-1} = a_j \dots a_{j+k-1}$, where $1 \leq i \leq m-k+1$ and $1 \leq j \leq n-k+1$. Then $o_s^{dir} = o(s) + \mathrm{Mass}(b_1 \dots b_{i-1}) - \mathrm{Mass}(a_1 \dots a_{j-1})$. The multiplicity $\mu(o_s^{dir})$ of o_s^{dir} is calculated as the number of tags from $\mathcal{T}(s)$ that match **a**; note that it may be smaller than $\mu(s) = |\mathcal{T}(s)|$.

Similarly, each string $s' \in \mathcal{J}_\mathbf{a}^{rev}$ imparts **a** a *reversed offset* associated with its end, which can be derived along with its multiplicity by applying the above procedure to $\overline{s'}$ and $\overline{s_\mathbf{a}}$ appropriately aligned against each other.

In this way, an aggregated string **a** gets assigned two sets of offsets—those of direct and reversed ones; note that one of the two may be empty.

Scaled and binned offsets. To be able to manage the offsets more efficiently, we transform them into *scaled offsets*, which amounts to multiplying each offset by a constant of the form 10^h for a sufficiently large h and rounding the obtained value to the nearest integer (in our experiments, $h = 4$). A direct (resp., reversed) scaled offset o^s has multiplicity $\mu(o^s)$ that can be calculated as the sum of multiplicities of the original direct (resp., reversed) offsets that got transformed into o^s. In addition, we maintain *binned offsets* representing integers. For an offset o, its corresponding binned offset o^b is obtained by rounding o to the nearest integer.

Consider a direct binned offset o^b; let $o_1, \dots, o_g$ denote its respective original direct offsets. We maintain for o^b the list of direct scaled offsets $o_1^s, \dots, o_g^s$ corresponding to $o_1, \dots, o_g$, respectively, along with their multiplicities. The multiplicity $\mu(o^b)$ of o^b is calculated as the sum $\mu(o_1^s) + \dots + \mu(o_g^s)$. Reversed binned offsets are treated in a similar way.

Finally, we define the main direct and reversed offsets of **a** in the following way. To obtain the *main direct offset*, or simply *offset*, $\mathbf{o}_\mathbf{a}^{dir}$ of an aggregated string **a**, we pick up the direct binned offset o^b of **a** of the highest multiplicity, then select its corresponding scaled offset o^s of the highest multiplicity, and let $\mathbf{o}_\mathbf{a}^{dir} = o^s \cdot 10^{-h}$. The *main reversed offset*, or simply *reversed offset*, $\mathbf{o}_\mathbf{a}^{rev}$ is defined analogously.

Protein mass estimation. For an aggregated string **a** that has both direct and reversed offsets, we define the *precursor mass estimate (PME) induced by* **a** as $M_\mathbf{a} = \mathbf{o}_\mathbf{a}^{dir} + m(s_\mathbf{a}) + \mathbf{o}_a^{rev}$; the score of the latter is set to $\mu(\mathbf{o}_\mathbf{a}^{dir}) \cdot \mu(\mathbf{o}_\mathbf{a}^{rev})$. A high score of $M_\mathbf{a}$ typically witnesses correctness of **a**; in addition, repetitive PMEs likely correspond to the intact masses of sufficiently well-represented proteins from the sample being analyzed, possibly upon adjustment based on expected PTMs.

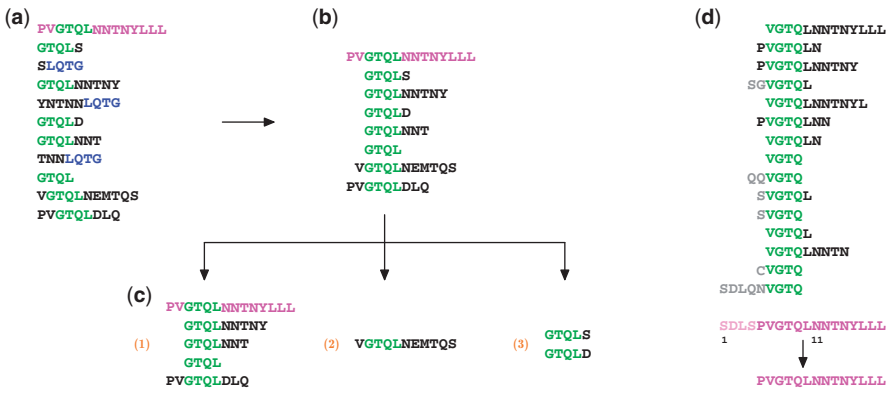

Fig. 3. (a–c) Processing the alemtuzumab dataset: handling the 4-mer 'GTQL' with the score 8. (**a**) The set of active strings containing 'GTQL' or 'LQTG'. (**b**) The aligned strings (duplications due to string reversal are not shown). (**c**) Clusterization: the string 'VGTQLNEMTQS' did not match well the main string 'PVGTQLNNTNYLLL', and gave rise to the second cluster; each of the strings 'GTQLS' and 'GTQLD' produced one mismatch when being aligned against 'PVGTQLNNTNYLLL' or 'VGTQLNEMTQS', and thus, both were placed in the third cluster. (**d**) The aggregated string with the amino acid sequence 'PVGTQLNNTNYLLL' (at the bottom) was obtained at time of processing the 4-mer 'VGTQ' with the score 25. The first four amino acids of its underlying consensus string 'SDLSPVGTQLNNTNYLLL' were eliminated by the validation procedure. Incorrect candidate amino acids in the aligned strings are marked gray (the amino acid scores are not shown). A 4-mer being processed and its reversed copy are highlighted

3 Results

We implemented our method in Java and benchmarked it on the top-down datasets for the light chain of alemtuzumab and CAH2 published in (Vyatkina *et al.*, 2015). In either case, the input spectra were deconvoluted using MS-Deconv (Liu *et al.*, 2010), and a list $\mathcal{N}$ of *de novo* amino acid strings was generated with the Twister algorithm from (Vyatkina *et al.*, 2015) using the default parameters (tag length: 4; mass tolerance: 4 mDa; peak reflection applied to individual deconvoluted spectra, and water loss ions eliminated prior to tag generation). The resulting aggregated strings for the CAH2 and alemuzumab datasets are listed in the supplementary file Aggregated-strings.xls.

3.1 Parameter selection

The proposed method depends on two parameters: the tag length k and the threshold *MinScore* on the score of a k-mer to be processed (see Supplementary Algorithm 1). The former needs to be the same as used for *de novo* strings generation following the approach from (Vyatkina *et al.*, 2015); a reasonable choice of k should imply that duplications are unlikely to occur among the tag strings and their reversed copies. The experimental results we report on were obtained for the tag length $k = 4$.

Observe that if *MinScore* is set to 1, each *de novo* string from $\mathcal{N}$ will contribute to some aggregated string from $\mathcal{A}$. In our experiments, we let *MinScore* = 2, so that the 'least confirmed' *de novo* strings would be left out.

3.2 CAH2

From the CAH2 dataset, 70 aggregated strings were derived. The ones that induce high-scoring PMEs represent or incorporate long sequence fragments of either CAH2 or a contaminant protein. The native—i.e. coming from bovine erythrocytes—contaminants comprise ubiquitin, phosphatidylethanolamine-binding protein, flavin reductase, cytochrome b5, hemoglobin alpha chain, ribonuclease UK114-like, Cu/Zn superoxide dismutase, and thymosin beta 4-like, and the extraneous ones—50S ribosomal protein L7/L12, 50S ribosomal protein L9, and cold-shock protein GroES from Shewanella, alcohol dehydrogenase from *S.cerevisiae*, and a mouse cytochrome c oxidase subunit 7B. All the identifications could be made by searching long enough amino acid sequences of the obtained aggregated strings against the non-redundant database using BLAST (Altschul *et al.*, 1990), except for Cu/Zn superoxide dismutase and thymosin beta 4-like, identification of which was sustained by the analysis of the initial *de novo* strings described in detail in (Vyatkina *et al.*, 2015). In particular, the extraneous contaminants could be immediately attributed to the experiments previously carried out on the same instrument. Subsequently, shorter amino acid sequences were matched against the CAH2 and contaminant sequences to decide on their originality.

The sequence coverage of CAH2 with the matching fragments of the aggregated strings, as compared to that induced by the initial *de novo* strings, decreased from 173 to 146 out of 260 amino acids (i.e. from 66.54 to 56.15%); see Supplementary Figure 1a. The 50-aa long gap in the coverage—from P-137 to N-186—hints that most spectra coming from CAH2 were actually acquired from its fragments rather than the entire protein. This is consistent with the fact that each of the PMEs brought forth by the aggregated strings originating from CAH2 is substantially smaller than the theoretical mass of 29 095.717 Da of the latter, except for the one of 45 846.421 Da induced by the 3rd aggregated string. In particular, the main offsets and PMEs induced by the 1st, 4th, 6th, 18th, 34th and 57th aggregated strings attributed to CAH2 immediately suggest that several underlying spectra of those were due to its C-terminal peptide 'PNVLDYWTYPGSLTTPPLLESVTWIVLKEPISVSSQQM LKFRTLNFNAEGEPELLMLANWRPAQPLKNRQVRGFPK,' N-terminal peptides 'MSHHWGYGKHNGPEHWHKDFPIANGER QSPVDIDTKAVVQDPALKPLALVYGEATSR' and 'MSHHW GYGKHNGPEHWHKDFPIANGERQSPVDIDTKAVVQDPALKP LAL' (upon N-terminal methionine truncation and serine acetylation) and internal peptides 'GEATSRRMVNNGHSFNVEYD DSQDKAVLKDGPLTGTY,' 'RLVQFHFHWGSSDDQGSEHTV DRKK,' and 'RLVQFHFHWGSSDDQGSEHTVDRKKYAAELHL VHWNTK,' respectively. However, this assumption should be treated with caution, since the two main offsets of an aggregated string may be defined by two distinct sets of spectra. This is illustrated by the 2nd aggregated string, the main direct and reversed offset of which was derived from CID/HCD and ETD spectra, respectively; as a consequence, the resulting PME exceeds by approximately 17.023 Da the theoretical mass of the

peptide 'MSHHWGYGKHNGPEHWHKDFPIANGERQSPVDI DTK' matched by the main offsets (again assuming N-terminal methionine truncation and serine acetylation, which caused the amino acid 'E' with the mass equal to that of acetylated serine to show up instead of the dimer 'MS' in the 2nd aggregated string). Thus, a more accurate interpretation would be to say that peptides starting and ending at the same locations in the CAH2 sequence as the above-listed ones must have appeared in the sample as degraded products.

Note that among the direct offsets with the second-highest multiplicity of the 3rd aggregated string, there is a one of 25 399.735 Da, which together with the main reversed offset of 1735.007 Da would lead to the PME of 29 005.646 Da differing by approximately 1Da from the theoretical mass of 29 006.687 of CAH2 upon N-terminal methionine truncation and serine acetylation (this discrepancy can be attributed to ±1 Da errors that commonly occur during deconvolution). Here, the confusion between the wrong and correct offset value is partially due to low multiplicities of both candidates: 10 and 5, respectively. However, a more thorough analysis could allow for selecting the correct one. First, the (incorrect) direct offset of 42 240.511 Da is due to a single spectrum (HCD, SCANS = 2318), for which the precursor charge was determined at time of deconvolution as 57. This abnormally large charge state suggests there is an error both in the precursor charge and precursor mass for this spectrum, and the latter leads to incorrect offsets associated with the tags determined by reflected counterparts of the original peaks, which give rise to the direct offset being examined. Second, the correct offset of 25 399.735 Da is supported by 4-tags derived from two distinct spectra (HCD, SCANS = 2326 and 2329), while any other candidate offset is sustained by tags from a single spectrum.

The precursor mass estimates of 14 174.47, 14 174.493 and 14 174.454 Da induced by the 16th, 32th and 40th aggregated strings, respectively, closely match the theoretical mass of 14 174.491 Da of ribonuclease UK114-like, to which they are attributed, upon N-terminal methionine truncation and serine acetylation. In particular, this validates identification of the latter two aggregated strings, either being too short on its own to allow for a meaningful analysis via BLAST search. On the other hand, these observations together provide an evidence of the presence in the sample of ribonuclease UK114-like modified as stated above.

Similarly, the 8th, 9th and 30th aggregated strings produce the PMEs of 8575.626, 8575.61 and 8575.631 Da, respectively, which accurately match the theoretical mass of 8575.616 Da of bovine ubiquitin upon N-terminal methionine oxidation, and thus, indicate its presence in the sample. At the same time, this certifies the origin of the short, and not fully correct, 30th aggregated string. Moreover, the 5th aggregated string gives the PME of 8559.618 Da, which closely approximates the mass of 8559.621 Da of unmodified ubiquitin, implying this form is observed as well. Interestingly, its direct offset with the second-highest multiplicity, equal to 5273.847 Da, leads to the PME of 8289.466 Da, and points to yet another form of ubiquitin—and namely, the one without the C-terminal 'RGG', the theoretical mass of which is 8289.477 Da.

3.3 Fab region of alemtuzumab

From the alemtuzumab dataset, 92 aggregated strings were generated. Through a BLAST search of their amino acid sequences against the non-redundant database, two contaminants could be identified, and namely, 50S ribosomal protein L29 from *Synechococcus* sp. PCC 7002, and 30S ribosomal protein S17 from Synechococcus. Moreover, based on the analysis of the original *de novo* strings presented in (Vyatkina *et al.*, 2015), we attributed three aggregated strings to a mouse Amy1 protein. As in the case of CAH2, the extraneous contaminants were due to the previous experiments.

The matching fragments of the *de novo* strings together covered 177 (82.71%) out of 214 amino acids of the light chain, and only 7 amino acids were lost when we switched to the aggregated strings, which thus led to the coverage of 170 (79.44%) amino acids (Supplementary Figure S1b). Not surprisingly, we could also obtain an accurate PME for this protein: the PMEs of 23 554.87, 23 553.861 and 23 553.876 Da induced by the 1st, 6th and 9th aggregated string, respectively, closely match the theoretical mass of 23 556.703 Da of the light chain of alemtuzumab, while the discrepancies can be attributed to ±1 Da errors at time of deconvolution, and imperfect reduction (the contribution of one remaining disulphide bond would be approximately –2.016 Da).

On the contrary, the coverage of the Fd region of the heavy chain with the matching fragments of the *de novo* strings was modest, comprising only 157 (68.86%) out of 228 amino acids, and upon generation of the aggregated strings, it was reduced down to 112 (49.12%) amino acids (Supplementary Figure S1c). Similarly to the case of CAH2, long gaps occur in the coverage, again meaning that many spectra originated from sequence fragments rather than the intact protein, and this is consistent with the PMEs reported by Twister. For example, the 4th and 5th aggregated string attributed to this protein induce the PME of 6645.354 and 6645.367 Da, respectively, thus indicating that several underlying spectra were acquired from its N-terminal peptide 'QVQLQESGPGLVRPSQTL SLTCTVSGFTFTDFYMNWVRQPPGRGLEWIGFIRDKAKGYT' with a pyroglutamic acid instead of N-terminal glutamine.

Finally, the 11th and 47th aggregated string gives rise to the PME of 7874.359 and 7874.351 Da, respectively, which accurately approximate the theoretical mass of 7875.344 Da of 50S ribosomal protein L29 from Synechococcus upon truncation of N-terminal methionine and oxidation of the second methionine at position 57, assuming errors of –1Da occurred during deconvolution.

4 Discussion

We have introduced a method for combining *de novo* sequence fragments, retrieved from deconvoluted top-down MS/MS spectra using our approach described in (Vyatkina *et al.*, 2015), into the aggregated strings endowed with direct and reversed offsets. This gives us longer parts of sequences of proteins sufficiently well-represented in the sample, and helps to promptly focus our attention primarily on those.

The main offsets of the aggregated strings often appropriately reflect their location in the sequence of the respective protein, and can serve to obtain an accurate intact mass estimate of a target protein, decide on which out of a few candidate proteoforms is contained in the sample, or facilitate characterization of the PTMs present in the protein being analyzed. We emphasize that the second and third tasks can be accomplished even without explicit evidence in the input spectra of the amino acids that differ between distinct proteoforms, or of the respective PTMs. Importantly, thereby an aggregated string with a very short, and not informative on its own, amino acid sequence can also be of true value.

On the other hand, the input MS/MS data may sometimes allow to quickly and unambiguously locate a PTM detected in this way. The proposed method can be further adjusted so as to solve this problem as well.

A careful analysis of the entire sets of direct and reversed offsets associated with the obtained aggregated strings can potentially

provide further insight into the composition of the sample; we consider development of algorithms for this purpose as one of the subsequent tasks to be addressed. Alternatively, one can manually browse through the input *de novo* strings—each with a single associated offset—that correlate well with the aggregated strings of interest, to gain additional information about the respective proteins.

Our present work constitutes a next step towards complete *de novo* sequencing of proteins from top-down spectra *alone*, for which purpose, to the best of our knowledge, no methods exist so far. While the current limitations of the top-down technology in general would prevent the proposed approach from retrieving the *entire* protein sequence, since the latter is typically not fully covered even with short tags, the aggregated strings can undoubtedly serve as an excellent base for the full reconstruction. However, achieving this goal will require more algorithmic developments, thus providing yet another promising direction for future research.

Funding

This work was supported by Government of Russian Federation [11.G34.31.0018 to K.V., P.A.P. until December 2014], Russian Science Foundation [14-50-00069 to K.V., P.A.P., since February 2015] and the Netherlands Organization for Scientific Research [93511034 to L.J.M.D., M.M.D.].

Conflict of Interest: none declared.

References

Altschul,S.F. *et al*. (1990) Basic local alignment search tool. *J. Mol. Biol*, **215**, 403–410.

Bandeira,N. *et al*. (2004) Shotgun protein sequencing by tandem mass spectra assembly. *Anal. Chem*., **76**, 7221–7233.

Bandeira,N. *et al*. (2007) Shotgun protein sequencing: assembly of peptide tandem mass spectra from mixtures of modified proteins. *Mol. Cell. Proteomics*, **6**, 1123–1134.

Bandeira,N. *et al*. (2008) Automated de novo protein sequencing of monoclonal antibodies. *Nat. Biotechnol*., **26**, 1336–1338.

Bertsch,A. *et al*. (2009) *De novo* peptide sequencing by tandem MS using complementary CID and electron transfer dissociation. *Electrophoresis*, **30**, 3736–3747.

Bondarenko,P.V. *et al*. (2009) Mass measurement and top-down HPLC/MS analysis of intact monoclonal antibodies on a hybrid linear Quadrupole Ion TrapOrbitrap Mass Spectrometer. *J. Am. Soc. Mass Spectrom*., **20**, 1415–1424.

Castellana,N.E. *et al*. (2010) Template proteogenomics: sequencing whole proteins using an imperfect database. *Mol. Cell. Proteomics*, **9**, 1260–1270.

Chi,H. *et al*. (2010) pNovo: De novo peptide sequencing and identification using HCD spectra. *J. Proteome Res*., **9**, 2713–2724.

Chi,H. *et al*. (2013) pNovo+: de novo peptide sequencing using complementary HCD and ETD tandem mass spectra. *J. Proteome Res*., **12**, 615–625.

Dancik,V. *et al*. (1999) De novo peptide sequencing via tandem mass spectrometry. *J. Comput. Biol*., **6**, 327–342.

Datta,R. and Bern,M. (2008) Spectrum fusion: using multiple mass spectra for de novo peptide sequencing. In: Vingron,M. and Wong,L. (eds.) *Research in Computational Molecular Biology, Lect. Notes Comput. Sci*., vol. **4955**, Springer, Berlin, Heidelberg, pp. 140–153.

Frank,A. and Pevzner,P. (2005) Pepnovo: de novo peptide sequencing via probabilistic network modeling. *Anal. Chem*., **77**, 964–973.

Fornelli,L. *et al*. (2012) Analysis of intact monoclonal antibody IgG1 by electron transfer dissociation Orbitrap FTMS. *Mol. Cell. Proteomics*, **11**, 1758–1767.

Guthals,A. *et al*. (2013) Sequencing-grade de novo analysis of MS/MS triplets (CID/HCD/ETD) from overlapping peptides. *J. Proteome Res*., **12**, 2846–2857.

He,L. and Ma,B. (2010) Adepts: advanced peptide de novo sequencing with a pair of tandem mass spectra. *J. Bioinf. Comput. Biol*., **8**, 981–994.

Horn,D.M. *et al*. (2000) Automated de novo sequencing of proteins by tandem high-resolution mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A*., **97**, 10313–10317.

Kellie,J.F. *et al*. (2010) The emerging process of Top Down mass spectrometry for protein analysis: biomarkers, protein-therapeutics, and achieving high throughput. *Mol. Biosyst*., **6**, 1532–1539.

Liu,X. *et al*. (2009) Automated protein (re)sequencing with MS/MS and a homologous database yields almost full coverage and accuracy. *Bioinformatics*, **25**, 2174–2180.

Liu,X. *et al*. (2010) Deconvolution and database search of complex tandem mass spectra of intact proteins: A combinatorial approach. *Mol. Cell. Proteomics*, **9**, 2772–2782.

Liu,X. *et al*. (2014) *De novo* protein sequencing by combining top-down and bottom-up tandem mass spectra. *J. Proteome Res*., **13**, 3241–3248.

Ma,B. *et al*. (2003) Peaks: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom*, **17**, 2337–2342.

Ma,B. (2015) Novor: real-time peptide de novo sequencing software. *J. Am. Soc. Mass Spectrom*., **26**, 1885–1894.

Mann,M. and Wilm,M. (1994) Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem*., **66**, 4390–4399.

Pevzner,P. *et al*. (2004) De novo repeat classification and fragment assembly. *Genome Res*., **14**, 1786–1796.

Robotham,S.A. *et al*. (2016) UVnovo: a de novo sequencing algorithm using single series of fragment ions via chromophore tagging and 351 nm ultraviolet photodissociation mass spectrometry. *Anal. Chem*, doi:10.1021/acs.analchem.6b00261.

Savitski,M. *et al*. (2005) New data base-independent, sequence tag-based scoring of peptide MS/MS data validates mowse scores, recovers below threshold data, singles out modified peptides, and assesses the quality of MS/MS techniques. *Mol. Cell. Proteomics*, **4**, 1180–1188.

Taylor,J.A. and Johnson,R.S. (1997) Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom*., **11**, 1067–1075.

Vyatkina,K. *et al*. (2015) *De novo* sequencing of peptides from top-down tandem mass spectra. *J. Proteome Res*., **14**, 4450–4462.