De Novo Sequencing of Peptides from High-Resolution Bottom-Up Tandem Mass Spectra Using Top-Down Intended Methods

Kira Vyatkina[1,2,3,4,], Lennard J. M. Dekker[5], Si Wu[6], Martijn M. VanDuijn[5], Xiaowen Liu[7,8], Nikola Tolić[9], Theo M. Luider[5] and Ljiljana Paša-Tolić[9]*

[1] Center for Algorithmic Biotechnology, Institute of Translational Biomedicine, Saint Petersburg State University, 7-9 Universitetskaya nab., Saint Petersburg 199034, Russia

[2] Department of Mathematical and Information Technologies, Saint Petersburg Academic University, Russian Academy of Sciences, 8/3 Khlopina Str, Saint Petersburg 194021, Russia

[3] Department of Information Technologies and Programming, ITMO University, 49 Kronverksky pr., Saint Petersburg 197101, Russia

[4] Department of Computer Technologies and Informatics, Saint Petersburg Electrotechnical University "LETI", 5 ul. Professora Popova, Saint Petersburg 197376, Russia

[5] Department of Neurology, Erasmus University Medical Center, Postbus 2040, 3000 CA Rotterdam, The Netherlands

[6] Department of Chemistry and Biochemistry, University of Oklahoma, 101 Stephenson Pkwy, Norman, Oklahoma 73019, United States

[7] Department of BioHealth Informatics, Indiana University-Purdue University Indianapolis, 535 West Michigan Street, IT 475, Indianapolis, Indiana 46202, United States

[8] Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, 410 West 10th Street, Suite 5000, Indianapolis, Indiana 46202, United States

[9] Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, Washington 99352, United States

Correspondence

Kira Vyatkina

Leading Researcher, Center for Algorithmic Biotechnology, Institute of Translational Biomedicine, Saint Petersburg State University

7-9 Universitetskaya nab., Saint Petersburg 199034, Russia

E-mail: k.vyatkina@spbu.ru

Received: 07 31, 2016; Revised: 09 15, 2017; Accepted: 10 09, 2017

Abbreviations

UV: ultraviolet

HCD: higher-energy C-trap dissociation

AGC: automatic gain control

NCE: normalized collisional energy

CAH2: carbonic anhydrase 2

Keywords: bottom-up proteomics, high-resolution mass spectrometry, *de novo* sequencing.

Total number of words: 7643

Abstract

Despite high-resolution mass spectrometers are becoming accessible for more and more laboratories, tandem (MS/MS) mass spectra are still often collected at a low resolution. And even if acquired at a high resolution, software tools used for their processing do not tend to benefit from that in full, and an ability to specify a relative mass tolerance in this case often remains the only feature the respective algorithms take advantage of. We argue that a more efficient way to analyze high-resolution MS/MS spectra should be with methods more explicitly accounting for the precision level, and sustain this claim through demonstrating that a *de novo* sequencing framework originally developed for (high-resolution) top-down MS/MS data is perfectly suitable for processing high-resolution bottom-up datasets, even though a top-down like deconvolution performed as the first step will leave in many spectra at most a few peaks.

Significance of the study

De novo sequencing is the only option for an exhaustive analysis of proteins from an organism with an unknown genome, novel splice variants, and antibodies. Of particular importance for such methods is accuracy; however, the numerous existing algorithms for sequencing peptides from bottom-up spectra often fail to distinguish between the correct and alternative incorrect candidate sequences. In this work, we demonstrate that the Twister approach initially designed for *de novo* sequencing of polypeptides from top-down data can efficiently process sets of high-resolution bottom-up tandem mass spectra, and strongly outperforms the previously developed methods in terms of both accuracy and speed. We believe our results will stimulate a much wider adaptation of *de novo* sequencing as an effective and reliable means for analysis of peptides and proteins.

1 Introduction

Tandem mass spectrometry is a powerful tool for analyzing proteins and peptides, which is being rapidly developed further. Along with the classical bottom-up technology [1] comprising enzymatic digestion as the first stage, the more recently emerged top-down technique [2] analyzing intact proteins is steadily gaining popularity. Another important breakthrough was the introduction of Orbitrap instruments [3], which made high-resolution and high mass accuracy mass spectrometry affordable to many laboratories and clinics.

However, bottom-up tandem (MS/MS) mass spectra are still often collected at a low resolution, primarily to achieve higher speed and sensitivity. And when high resolution is acquired, the bottom-up data is typically processed with the software tools originally designed for low-resolution bottom-up MS/MS spectra, in which case selection of a smaller mass tolerance remains the only adaptation in the data processing for the high-resolution MS/MS data.

Among the essential tasks of mass spectrometry-based proteomics is *de novo* sequencing of peptides and proteins. Since a good mass spectrum acquired from a relatively short (e.g. tryptic) peptide can be expected to contain enough information for retrieving the entire amino acid sequence of the latter, most of widely used bottom-up *de novo* sequencing tools, including PEAKS [4], PepNovo [5], pNovo [6], Lutefisk [7], Sherenga [8], Vonode [9], and Novor [10], seek to interpret the input spectra on the individual basis.

Even though *de novo* sequencing can strongly benefit from high resolution and high mass accuracy, out of those seven methods, only two explicitly account for the precision of such data – and namely, Vonode originally intended for analyzing high-resolution datasets, and PepNovo, which was fine-tuned for this case [11]. But even for these two approaches, the accuracy of the output peptide sequences is still far below the results typically obtained for a database search, since in the latter case, only known sequences are brought into consideration, which makes dramatically smaller the number of sequence permutations potentially examined. So at the moment the *de novo* sequencing strategy is only considered when a database is not available.

To the best of our knowledge, the only top-down *de novo* sequencing method is the recently proposed Twister [12-14]. Typically, a top-down MS/MS spectrum provides modest sequence coverage of the underlying protein, and this is taken into consideration by the Twister approach, which extracts from a spectrum short but notably accurate sequence fragments, and further combines those across several spectra supposed to map to the same protein or peptide. Prior to be passed to Twister, the input spectra should be deisotoped and charge state deconvoluted with MS-Deconv [15], or Xtract, or another top-down deconvolution software tool. Subsequently, we will refer to the entire procedure simply as *deconvolution*. It should be noted that though deconvolution of bottom-up MS/MS spectra also proved to be useful [11,16-19], many methods neither require nor perform it as a preprocessing step – again, unlike in the top-down case, in which it is usually indispensable.

Although conceptually, deconvolution of bottom-up and top-down spectra is just the same procedure, from the computational point of view, these two tasks are inherently different. In the former case, isotopic envelopes are small, and for the fragment ions with the masses below 1500 Da, the monoisotopic peak is typically the most intense, also called the *base peak*, and clearly dominating one. In the latter case, the envelopes are mostly large, with the base peak somewhere in the middle, and the monoisotopic peak either missing or indistinguishable from noise, which leads, in particular, to the notorious errors by ±1 Da in the deconvoluted masses due to an incorrect assignment of the number of heavy isotopes of carbon that contribute to the base peak [20]. Moreover, in a top-down spectrum, isotopic envelopes can overlap in a fairly complex manner, which further complicates its processing. Quite naturally, the strategies employed by bottom-up and top-down deconvolution approaches also differ from each other: while the former seek to detect and eliminate small groups of isotopic peaks belonging to a same cluster, the latter generally eliminate all the peaks that failed to be assigned to any "good" envelope.

While the top-down methods Z score [21] and THRASH [22] were tested on MS1 spectra from peptide digests, we are unaware of any mention in the literature of an attempt to deconvolute bottom-up MS/MS spectra with a top-down method or vice versa, except for an unsuccessful application of MS-Deconv to high-resolution bottom-up MS/MS spectra that led to poor results reported in [23].

The goal of the present work is to demonstrate that top-down deconvolution algorithms *are* applicable to high-resolution bottom-up MS/MS spectra, provided that the deconvoluted spectra are further processed in a top-down like fashion. In the majority of those, at most a few peaks are left, making their individual interpretation no longer possible; however, the remaining peaks often define highly accurate peptide sequence tags. Being capable of generating such tags is important on its own, due to their wide usage as filters in the context of database search [7,16,24-33]. Moreover, this implies suitability of the Twister concept to deconvoluted high-resolution bottom-up MS/MS data, which results in an unprecedentedly accurate *de novo* reconstruction of sequence fragments of the peptides contained in the sample.

To sustain our arguments, we closely examine behavior of MS-Deconv on a bottom-up dataset acquired at a high resolution from carbonic anhydrase 2 (CAH2), assess the quality of the derived sequence tags, evaluate performance on it of the Twister approach coupled to MS-Deconv or Xtract, or the bottom-up deconvolution tool Mascot Distiller [34], and compare the *de novo* sequencing results to those produced by PepNovo, Vonode and Novor. In addition, we illustrate applicability of Twister to analyzing simple protein mixtures through discussing its behavior on a mixture of 5 proteins and benchmarking it against the other methods. The version 2.2.1 of Twister we used is freely available at http://bioinf.spbau.ru/en/twister, along with the test datasets and sample results. The only distinction between the version 2.2.1 (bottom-up and top-down) and 2.2 (solely top-down) of Twister consists in the ability of the former to recognize carbamidomethylated cysteines.

2 Methods

2.1 Dataset acquisition

CAH2: main experiment

CAH2 was purchased from Sigma-Aldrich (St. Louis, MO). CAH2 (1 pmol) was dissolved, reduced with dithiothreitol (DTT), alkylated with iodoacetamide, and digested overnight with trypsin, GluC or Lys-C; all enzymes were used in 1:50 (w/w) ratio. All LC-MS measurements were carried out on a nano LC system (Ultimate 3000; Thermo Fisher Scientific, Gemeringen, Germany) online coupled to a Q-Exactive plus MS (Thermo Fisher Scientific, Bremen, Germany). One microliter of digested sample was injected onto the nano LC system, with a C18 trap column (PepMap C18, 300 µm ID × 5 mm, 5 µm particle size and 100 Å pore size; Thermo Fisher) and a 50 cm analytic column (PepMap C18, 75 µm ID × 500 mm, 2 µm particle size and 100 Å pore size; Thermo Fisher Scientific). 0.1% (v/v) formic acid in water was delivered at 250 nL/min, and peptides were eluted with a 30 minute linear gradient that reached 30.4% (v/v) acetonitrile in 0.1% formic acid. All solvents used were purchased from Biosolve (Valkenswaard, Netherlands). The separation of the peptides was monitored by a UV detector (absorption at 214 nm).

High resolution full scan MS data was obtained (resolution 70000, AGC 1E^6, max. injection time 100 ms, mass range of *m*/*z* 375.00-1500.00), MS/MS spectra were obtained by HCD fragmentation applying 30% NCE. MS/MS was performed on the top ten most intense mono isotopic masses (charge 2-5) in the full scan spectra (resolution 17500, AGC 5E^5, max. injection time 50 ms) and no dynamic exclusion. In total, 177741 HCD MS/MS spectra were acquired (trypsin: 91747 spectra, GluC: 43026 spectra, Lys-C: 42968 spectra). In our experiments, we also separately used the subset of 47536 tryptic MS/MS spectra contained in the files 140411_QE_Cah-1.raw, 140411_QE_Cah-2.raw, 140411_QE_Cah-3.raw, and 140411_QE_Cah-4.raw; in what follows, this subset and the entire set of spectra is referred to as “tryptic” and “full” dataset, respectively.

5-protein mixture

Human serum albumin, chicken egg albumin, horse myoglobin, bovine catalase, and bovine carbonic anhydrase 2 (CAH2) were purchased from Sigma-Aldrich (St. Louis, MO). For each protein, its solution at a concentration of 2 mg/mL was prepared, and the solutions were mixed in equal amount. The total amount of sample analyzed was 500 ng. The protein mixture was reduced with dithiothreitol (DTT), alkylated with iodoacetamide, and digested overnight with trypsin used in 1:50 (w/w) ratio. All LC-MS measurements were carried out on a nano LC system (Ultimate 3000; Thermo Fisher Scientific, Gemeringen, Germany) online coupled to a Q-Exactive Orbirap HF (Thermo Fisher Scientific, Bremen, Germany). One microliter of digested sample was injected onto the nano LC system, with a C18 trap column (PepMap C18, 300 µm ID × 5 mm, 5 µm particle size and 100 Å pore size; Thermo Fisher) and a 50 cm analytic column (PepMap C18, 75 µm ID × 250 mm, 2 µm particle size and 100 Å pore size; Thermo Fisher Scientific). 0.1% (v/v) formic acid in water was delivered at 250 nL/min, and peptides were eluted with a 30 minute linear gradient that reached 30.4% (v/v) acetonitrile in 0.1% formic acid. All solvents used were purchased from Biosolve (Valkenswaard, Netherlands). The separation of the peptides was monitored by a UV detector (absorption at 214 nm).

High resolution full scan MS data was obtained (resolution 60000, AGC 3E^6, max. injection time 60 ms, mass range of *m*/*z* 375.00-1500.00), MS/MS spectra were obtained by HCD fragmentation applying 28% NCE. MS/MS was performed on the top twenty most intense mono isotopic masses (charge 2-5) in the full scan spectra (resolution 30000, AGC 5E^5, max. injection time 45 ms, isolation width 2.0) and no dynamic exclusion. In total, 27473 HCD MS/MS spectra were acquired.

CAH2: additional experiments

Sampe preparation was carried out precisely like in the case of 5 protein mixture. Additional experiments comprised MS/MS measurements for different values of isolation width and different amount of CAH2. Within the first group of experiments, high resolution full scan MS data was obtained (resolution 120000, AGC 4E^5, max. injection time 50 ms, mass range of *m*/*z* 375.00-

1500.00); MS/MS spectra were obtained by HCD fragmentation applying 28% NCE. MS/MS was performed on the top twenty most intense mono isotopic masses (charge 2-7) in the full scan spectra (resolution 15000, AGC $5E^5$, max. injection time 50 ms) for an isolation width of 0.5, 1, 1.5, 2, and 3; 35984, 36331, 35451, 36169 and 36039 MS/MS spectra were thereby acquired, respectively. For the experiments with the amount of sample, high resolution full scan MS data was obtained in the same way as in the above case; MS/MS spectra were obtained by HCD fragmentation applying 28% NCE. MS/MS was performed on the top twenty most intense mono isotopic masses (charge 2-7) in the full scan spectra applying an isolation width of 1.5; for the resolution of 15000, we used AGC $5E^5$ and max. injection time 50 ms; the amount of sample analyzed was 100 amol, 1 fmol, 10 fmol, 100 fmol and 1 pmol, respectively. Moreover, 1 fmol of sample was analyzed at a resolution of 30000 with max. injection time 45 ms, and of 60000 with max. injection time 118 ms, and AGC $2E^5$ in either case. In these seven experiments, 31193, 30477, 29225, 31752, 35067, 19687 and 10938 HCD MS/MS spectra were obtained, respectively.

All the raw data is available from http://bioinf.spbau.ru/en/twister.

2.2 Deconvolution

Twister takes as input a set of deconvoluted MS/MS spectra, in contrast to Vonode and PepNovo, which process the original ones. Throughout our experiments, deconvolution was accomplished using the top-down intended tools MS-Deconv [14] or Xtract, or the bottom-up intended tools Mascot Distiller [34] and Hardklör [35,36].

In case of MS-Deconv, the collected raw bottom-up MS/MS spectra were centroided and converted into mzXML format with ReAdW 4.3.1, and subsequently deconvoluted using the version 0.8.07370 of this software tool with the default parameters (maximum charge state: 30; maximum monoisotopic mass of fragment ions: 49000 Da; signal-to-noise ratio: 1; envelopes of precursor ions were deconvoluted to derive the precursor masses of MS/MS spectra). We also ran a few experiments with the values of the first two parameters closer to those typically applied when processing peptides; however, this led to no positive effect. Even the running time was not visibly improved, since MS-Deconv always proceeded within minutes. Conversely, bounding a charge state with a small value like 5 sometimes resulted in a lower-quality output. Moreover, we compared the sets of *de novo* strings produced by Twister upon deconvolution with the default parameters, and the first and second parameter set to 10 and 5000 Da, respectively: the results were almost identical for 3- and 4-tags, respectively, and fully identical for 5- and 6-tags, respectively. Therefore, our choice was to simply use the default parameter settings throughout the experiments described in detail below.

Thermo Xtract 3.0 was also run with its default top-down parameters, including, in particular, Low and High Mass of 56.00 and 60000.00, and signal-to-noise ratio of 2; the only exception was the maximum charge state, which we set to 10 (while the default value is 25). Since Xtract typically leaves more peaks in a spectrum, as compared to MS-Deconv, and moreover, works substantially slower than MS-Deconv, we chose to lower this value.

The parameters used for Mascot Distiller 2.5.1.0 are listed in Supplementary materials.

Hardklör was run with the default parameters, except for ‘resolution’, which was set to 17500 and 30000 for CAH2 and 5-protein mixture, respectively, and ‘centroided’ and ‘ms_level’ set to 1=yes and 2, respectively, in either case (the default values of those three parameters are 60000, 0=no, and 1, respectively).

2.3 The Twister approach

The input MS/MS spectra are first deconvoluted and preprocessed. The latter includes:

- merging of peaks with the same mass, up to a predefined tolerance ε (i.e. those deconvoluted from distinct charge states),

- addition of auxiliary peaks that for a peptide consisting of *n* amino acids mimic the (non-existing) b_0- and y_0-, and b_n- and y_n-ions needed for identification of the first and last amino acids (e.g. in case of an HCD spectrum, the corresponding extra peaks are introduced at the zero mass, the molecular mass of water *Mass*(H_2O), and at *PM*–*Mass*(H_2O) and *PM*, where *PM* denotes the precursor mass.), and

- optional *peak reflection* and *water loss ions elimination*.

Peak reflection amounts to generating for a peak *p* with the mass *M*(*p*) an artificial peak with the same intensity at the mass *PM*–*M*(*p*). Note that for a peak corresponding to a *b*- or *y*-ion, its reflected copy corresponds to the complementary *y*- or *b*-ion. In this way, we seek to generate peaks that would represent missing fragment ions, provided that their complementary ions are observed, thus potentially prolonging ladders of *b*- and *y*-ions and obtaining more *k*-tags. Water loss ion elimination is performed as follows: if for a peak *p* with the mass *M*(*p*), another peak *p'* with the mass *M*(*p*)-*Mass*(H_2O) is present in the spectrum, *p*' is supposed to appear due to a water loss from *p*; then the intensity of *p* gets increased by that of *p'*, and *p'* is eliminated.

In what follows, mass differences are always evaluated up to 2ε.

The core of the Twister approach [12] comprises three stages:

- generation of tags of length *k*, or *k*-tags, for a fixed *k*,
- *T*-Bruijn graph construction, and
- optimal paths computation and *de novo* strings extraction;

see Supplementary materials for a brief description of those, and [12] for further details.

Throughout our experiments, we ran Twister with the default parameters (tag length *k*=4, mass tolerance ε=4 mDa, peak reflection and water loss ion elimination applied to the individual deconvoluted spectra at the preprocessing stage), except for a few cases when either a different tag length was specified, or the adopted deconvolution software tool could combine input spectra not necessarily with the same precursor mass (namely, this holds for Xtract and Mascot Distiller), in which case peak reflection is inappropriate.

The results reported in this paper were generated using the version 2.2.1 of Twister.

3 Results

We benchmarked our approach on the high-resolution bottom-up dataset acquired from CAH2, and briefly examined its behavior on the one acquired from the 5-protein mixture.

3.1 Assessing the quality of *k*-tags

Deconvolution of the input spectra was performed using MS-Deconv. Subsequently, a set of *k*-tags was derived from the CAH2 dataset as described in Section 2.3, for *k* from 0 to 6. For the purpose of comparison, we generated two additional sets of *k*-tags, for *k* from 0 to 6, extracting those from a longest (instead of optimal) path in each connected component of the spectrum graph, and picking up all the possible *k*-tags, respectively.

Further, for the 4-tags from the main set, a *T*-Bruijn graph was constructed, from which a set N of *de novo* strings was computed (see the supplementary file `Twister_bottom-up_de-novo_CAH2.xls`; note that the *de novo* strings are listed by decreasing score calculated as explained in Supplementary materials). Having searched the long enough strings from N with BLAST [37] against the non-redundant database, again following [12], we detected and identified in the sample 27 native (bovine) and 5 extraneous contaminants, the latter being either enzymes used for digestion or human keratins (all the 32 contaminants are listed in Supplementary materials).

To form a set of spectra, from which a representative set of tags could be generated, we proceeded as follows. First, we searched the original spectra with MS-GF+ [38] against the database composed of CAH2 and the 32 contaminants using the parameters specified in Supplementary materials, selected the 19137 ones, for which the peptide-spectrum match reported by MS-GF+ had an E-value below 10^{-5}, and restricted our attention to their deconvoluted counterparts and the 463 identified peptides. Note that a search against the small database consisting of CAH2 and the contaminants identified based on the *de novo* strings, rather than e.g. UniProt, gave us identifications mostly for high-quality spectra, the deconvoluted counterparts of which contained tags (that contributed to our *de novo* strings), and therefore, the resulting set of spectra was particularly suitable for assessing tag quality,

Next, we filtered out 309 spectra with no peaks, 1729 spectra, for which the precursor mass reported by MS-Deconv was zero, and 8092 spectra, for which the latter did not meet the mass of the peptide matching a spectrum according to MS-GF+. The remaining 9007 spectra contained a total of 226274 peaks, for which the histogram in Figure 1 shows dependence of the number of peaks observed in isotope envelope derived by MS-Deconv on a peak mass; it witnesses, in particular, that MS-Deconv could successfully detect and process the isotopic envelopes consisting of as few as two peaks, even though its filtration functions implicitly assume the size of those is three or greater.

Statistics on the *k*-tags derived from those spectra before and after preprocessing, for *k* from 0 to 6, is summarized in the supplementary file `Twister_bottom-up_tag-statistics.xls`. Tags defined by *b*- and *y*-ions are referred to as *b*- and *y*-tags, respectively. *Shifted* tags are the ones defined by ladders of peaks with either the same neutral loss or the same error of ±1 Da introduced at time of deconvolution. Predominance of *y*-ions, and consequently, *y*-tags, in the initial deconvoluted spectra is consistent with what is typically expected of original HCD spectra [18,39]. Since often only one fragment ion (if any) from a pair of complementary ions is present in a spectrum, peak reflection at time of preprocessing increases the number of tags; besides, it equalizes the share of *b*-and *y*-ions and tags. Observe that 0-tags are simply peaks, and when considering all the possible 0-tags, we obtain statistics for fragment ions in the deconvoluted bottom-up spectra. When annotating tags, we used an error tolerance of 0.02 Da to validate the mass of each underlying peak – in contrast to the ultra-low error tolerance of 4 mDa applied to mass differences at time of introducing edges in spectrum graphs – since an absolute error in a peak mass still can be accordingly large.

We also report the percent of identified peptides, to which at least one annotated *k*-tag corresponds; we refer to them as being "hit". If a set of *k*-tags is used for database filtration, the peptides it hits will remain under consideration, as needed. Moreover, for *k* from 1 to 6, the coverage of a peptide sequence with the annotated *k*-tags is provided, averaged over either individual spectra or peptides. Similar values for 0-tags represent site cleavage statistics, and comprise 59.63% and 76.1% for individual spectra and peptides, respectively. These statistics indicate that a large fraction of a peptide sequence can often be assembled *de novo* from deconvoluted bottom-up spectra profiting from tags to form the seed of reconstruction.

The lists of *de novo* strings generated through construction of a *T*-Bruijn graph from the *k*-tags and extraction of optimal paths from its connected components, for *k* from 3 to 6, are provided in the supplementary file `Twister_bottom-up_de-novo_CAH2.xls`. For each *de novo* string, we check in which experiments the spectra were acquired that gave rise to its underlying tags, and list the respective enzymes. Note that tryptic and LysC-peptides can contribute to the same *de novo* string supported by *b*-ions if those peptides start at the same amino acid residue, or to the same *de novo* string supported by *y*-ions if the peptides end at the same residue. On the contrary, tags originating from GluC-peptides cannot be combined with those from tryptic ones or LysC-ones unless all the respective peptides are simultaneously N- or C-terminal.

To illustrate how the tags from different spectra can be combined, we provide the details on the formation of the 7th *de novo* string VLDALDSLK generated for CAH2 using 4-tags: in the supplementary file VLDALDSLK.xls, for each spectrum that gave rise to the 4-tags contributing to this string, we list the (longest possible) sequence tags, from which the respective 4-tags were extracted, and indicate the masses of their defining peaks.

3.2 Benchmarking against alternative approaches

As a next step, we sought to examine the possibility of using deconvolution tools other than MS-Deconv in combination with Twister, and compare efficiency of the Twister framework to that of the previously existing *de novo* sequencing methods. For the former purpose, we selected Thermo Xtract – widely applied to processing top-down data, and supplied by the instrument vendor, Mascot Distiller – a popular solution from Matrix Science for deconvoluting and processing bottom-up data, and Hardklör – a frequently used freely available tool. For the latter aim, our choice was PepNovo [5,11] and Vonode [9] adjusted and intended specifically for processing high-resolution bottom-up data, respectively, and Novor [10] originally benchmarked on low-resolution MS/MS data.

As Table 1 demonstrates, MS-Deconv filters out not firmly supported peaks most efficiently: the average number of peaks per spectrum was reduced by almost 19 and 15 times upon deconvolution of the full and tryptic dataset, respectively. For Xtract, the respective ratio is roughly 7, and for Mascot – approximately 1.5. In particular, these values immediately emphasize the difference between top-down and bottom-up intended deconvolution methods. For Hardklör, which is used for processing both bottom-up and top-down MS/MS spectra, this ratio is approximately 5.4.

While MS-Deconv and Hardklör spent only a few minutes per input file, Xtract required 1-2 hours to this end, and for Mascot Distiller, it took a few hours to process each file with tryptic spectra, and several hours would be needed for processing the files with GluC- and LysC-spectra. Therefore, when benchmarking MS-Deconv against Mascot Distiller ran with the parameters appropriate for high-resolution data, we restricted our attention to the tryptic dataset. We also used this dataset for comparing Twister to PepNovo and Novor, since PepNovo performs best on tryptic datasets, and Novor works only on those. Moreover, a comparison with Twister coupled to Hardklör was performed of the tryptic dataset, since from the full one, too many 4-tags were extracted for Twister to generate the *de novo* strings (for 5-tags this was possible, however). Comparison with Vonode was performed on the full dataset, as well as that with Twister coupled to Xtract. Whenever Twister was run, the underlying tag length was 4. Vonode was run with the default parameters. For PepNovo, the non-default parameters were the fragment and PM tolerance set to 0.01 and 0.02 Da, respectively, and PTMs specified as C+57 (cysteine carbamidomethylation). For Novor, the following parameters were used: enzyme=Trypsin (default), fragmentation=HCD, massAnalyzer=FT, fragmentIonErrorTol=0.01 Da, precursorErrorTol=10 ppm, fixedModifications=Carbamidomethyl (C). The parameter variableModifications was commented out; nevertherless, methionine oxidation was applied as such. The results are summarized in Table 2, and the corresponding *de novo* strings are listed by decreasing score in the supplementary file `Twister_bottom-up_de-novo_comparison.xls`. A *de novo* string or its fragment is *correct* if it fully matches the sequence of the target protein or a contaminant; for a string produced by Twister, a match between its reversed copy and protein sequence is also considered valid. The longest correct fragment of length at least 4 (if any) of each *de novo* string is highlighted in color, and the location of its first and last amino acid in the respective protein sequence is specified; if the former exceeds the latter, the fragment matches the sequence upon reversal.

It is visible at a glance that most accurate *de novo* strings are generated by the Twister approach from the spectra deconvoluted with MS-Deconv. In addition, the number of *de novo* strings produced by Twister is substantially smaller than the number of the strings output by Vonode, PepNovo or Novor (however, these three tools sequence individual spectra, while Twister efficiently combines tags extracted from different spectra). The correct fragments of the *de novo* strings derived from the full dataset by Twister launched after MS-Deconv cover 249 out of 260 amino acids composing the CAH2 sequence, which is 95.77%; see Figure 2. In particular, note that due to the N-terminal methionine oxidation and serine acetylation, the first two amino acids cannot be retrieved with the current version of Twister; on the other hand, the acetylated serine appears in some *de novo* strings as a glutamic acid, which has an identical mass.

However, since the strings obtained applying Twister together with Xtract, Mascot or Hardklör, or running Vonode, PepNovo or Novor, are essentially less accurate, a fairer way of comparing the efficiency of the approaches amounts to a comparison of the sequence coverage for CAH2 inferred by the entirely correct *de novo* strings. The fraction of such strings constitutes 61.8% and 62.07% in the output of Twister launched after MS-Deconv, while it is below 28% for PepNovo and much lower for other alternative approaches. In case of PepNovo, Vonode, and Novor, scores may help to distinguish between correct and wrong sequences; however, in the output of each of those tools, there are several incorrect amino acid strings even among the high-scoring ones. Moreover, the "incorrect"' strings reported by Twister often comprise only one or a few terminal spurious amino acids, while those produced by the other tools are often entirely wrong.

For the full dataset, the coverage of the CAH2 sequence with the entirely correct *de novo* strings produced by Twister from the spectra deconvoluted with MS-Deconv is the highest. In case of the tryptic dataset, the strings derived by Twister from the spectra deconvoluted with Mascot provide a slightly higher coverage (75.38% vs. 72.31%); however, this gain of 3% comes at an expense of a large fraction (almost 80%) of incorrect strings, several of which are long and high-scoring, and would be undistinguishable from the correct ones in the absence of *a priori* information on the target protein sequence. The best coverage (78.46%) is provided by Twister coupled to Hardklör, and the top-scoring *de novo* strings generated in this way are very accurate; however, less than 30% of all the strings are fully correct, and several of those are short and/or irrelevant.

Interestingly, the 436th *de novo* string SGGGYGGGSGSAEEGGGY generated by Twister from the spectra deconvoluted with Hardklör, being reversed, unambiguously pointed to yet another contamination protein – human cytokeratin 9.

3.3 Additional experiments

To gain a deeper understanding of benefits and shortcomings of the proposed pipeline, we carried out a few additional experiments. First, we examined the influence of the isolation width on the Twister results. Indeed, this matters, since top-down intended deconvolution methods eliminate fragment ions not assigned to "good" isotopic envelopes. Consequently, if the isolation width happens to be too small, there may be too few isotopic peaks in bottom-up MS/MS spectra, for those to still be informative upon top-down like deconvolution. We tested five settings for this parameters, and namely, 0.5, 1., 1.5, 2, and 3; all the other parameters (listed in the "Methods" section) were kept the same. The respective sets of the Twister *de novo* strings are provided in the supplementary file `TBruijn_bottom-up_de-novo_IW.xls`. It is visible at a glance that the results obtained for 1.5, 2, and 3 are quite similar; for 1 they become a bit worse, and for 0.5 more substantially deteriorate. However, it is most typical to acquire bottom-up MS/MS data applying isolation width between 1.5 and 3, which means that no special effort should be needed to make such data better suited for Twister from this point of view.

The second group of experiments was devoted to verifying, which amount of a sample may be sufficient for analyzing it with the Twister pipeline. First, we analyzed 100 amol, 1 fmol, 10 fmol, 100 fmol and 1 pmol of CAH2 at a resolution of 15000 (the corresponding lists of *de novo* strings are provided in the supplementary file `TBruijn_bottom-up_de-novo_amount.xls`, along with those from the subsequent experiments). It was immediately clear that under the respective experimental conditions (see the section "Methods" for details), at least 100 fmol is needed for getting meaningful results. Further, we increased the resolution to 30000 and next to 60000, thereby lowering the AGC threshold, and at the resolution of 60000, also increasing the maximum ion injection time. And under such settings, the results obtained for just 1 fmol of the sample were impressively better than those for 100 fmol and 1 pmol obtained at a resolution of 15000. This emphasizes again the importance of high resolution in the context of the proposed approach.

3.4 Analyzing protein mixtures

In order to evaluate behavior of Twister on simple protein mixtures, we carried out a bunch of computational experiments, similar to those outlined in Section 3.2, for a dataset acquired from the mixture of 5 proteins (human serum albumin, chicken egg albumin, horse myoglobin, horse catalase, and CAH2) as described above. To make this more precise, we processed this dataset with Twister coupled to MS-Deconv, Xtract or Hardklör, and with PepNovo, Vonode and Novor. All the software tools were launched with the same parameters as in the case of CAH2, except for the resolution, which was specified for Hardklör accordingly. The *de novo* strings derived in all the experiments are listed in the supplementary file `TBruijn_bottom-up_de-novo_5ProteinMix-comparison.xls`. The statistics on the total number of the *de novo* strings, the number and percentage of the correct ones, and the coverage of the five target protein sequences achieved by each tool is provided in the supplementary file `TBruijn_bottom-up_de-novo_5ProteinMix-coverage.xls` Note that Twister again generated long correct subsequences of each target protein. A number of long *de novo* strings did not match any target protein sequence; having searched them against the non-redundant database with BLAST, we detected and identified 11 contaminants (see Supplementary materials for a list of those), which once again testifies that Twister's output is unlikely to contain spurious sequences.

In general, the results obtained for the 5-protein mixture had much in common with those for CAH2. In particular, application of Twister following each of the tools MS-Deconv, Xtract and Hardklör led to similar coverage for human serum albumin, as well as for chicken egg albumin and horse myoglobin (however, Twister coupled to MS-Deconv performed a bit worse than the other two options on horse catalase and CAH2). But the percentage of the correct *de novo* strings was twice as high in the case of MS-Deconv as in the case of Xtract or Hardklör (38.88% vs 19.33 and 19.5%, respectively).

The percentage of the correct sequences generated by PepNovo was even slightly higher (and namely, 42.1%). However, this came at an expense of a substantially larger number of "not fully correct" strings (13591 vs 3656 reported by Twister launched after MS-Deconv). The sequence coverage inferred by Twister in conjunction with MS-Deconv and by PepNovo was approximately the same for each of human serum albumin, chicken egg albumin and CAH2, while Twister performed better than PepNovo for horse myoglobin and worse for horse catalase.

As in the case of CAH2, Vonode performed visibly worse in both "Sequence Tags" and "Consensus Sequence Tags" modes, and Novor eventually failed on this dataset.

Altogether, our observations indicate that Twister usually performs at least as good as the most competitive alternative approaches, while being notably faster, and producing an output of substantially smaller size, which makes the results remarkably easier to interpret.

4 Discussion

We have demonstrated that the Twister *de novo* sequencing framework designed for analyzing top-down data is perfectly suitable for processing bottom-up MS/MS spectra acquired on modern mass spectrometers at a high resolution. As an intermediate step, it generates a set of particularly accurate peptide sequence tags, which is affirmed by the statistics we gathered for the spectra from CAH2 with charge from 2 to 5 (as determined by MS-GF+). For the spectra with a smaller maximum charge, similar statistics will be even better.

We emphasize that the unannotated tags are not necessarily *incorrect*. In general, the presence in a spectrum of tags that do not conform to either a suggested interpretation of the former or each other should mean we are likely dealing with a mixed spectrum; see Supplementary materials for an example.

For the purpose of *de novo* sequencing, the most accurate tag generation from optimal paths seems the best, even though it misses some correct tags that can be derived from longest paths. However, in the context of tag-guided database search, consideration of all the available tags will likely be the choice of preference.

While top-down like deconvolution has proven itself to be applicable to high-resolution bottom-up spectra, it is also obvious that the respective software tools were not *intended* for this purpose. As an illustrative example, consider the first *de novo* string KVQAEGYVLALPKLAPDQVVAAVVQDPALKPLALVY generated from the CAH2 dataset at tag length 4. The underlined fragment matches the CAH2 sequence in reverse. It is easy to see that the subsequent fragment, not recognized as correct, represents a reversed copy of the suffix with length 15 of the underlined one. Essentially, its appearance was due to a few spectra, for which the precursor mass reported by MS-Deconv was twice as large as the true one; the respective peaks thus resulted from the reflection of the *y*-ions defining the tags that contributed to the correct fragment. While we have never observed a similar effect for top-down spectra, it sometimes manifests itself in the bottom-up case, and obviously requires a close examination. More generally, within our experiments, MS-Deconv failed to output the precursor mass for roughly 50% of spectra identified by MS-GF+ – i. e. high-quality ones, and Xtract performed even worse, again indicating there is much room for improvement, which represents an attractive future research direction. Another side effect to be eliminated are ±1 Da errors in masses as small as a few hundred daltons, which sometimes occur despite the fact that this should not happen in theory.

Top-down deconvolution methods play a fairly important role in the Twister pipeline. This becomes particularly evident if we compare the results that can be obtained with Twister coupled to such a method to those generated by Twister coupled to a powerful and popular bottom-up deconvolution tool, like Mascot Distiller or Hardklör. The latter processes MS/MS spectra substantially faster than Mascot Distiller but also eliminates only a modest fraction of peaks from those, which may lead to an excessive number of Twister-generated *k*-tags, most of which will be inevitably incorrect.

The Twister approach was extended to produce from the *de novo* strings the so-called aggregated strings endowed with offsets, which are supposed to correspond to the masses of the prefix and suffix of the protein sequence preceding and following the retrieved fragment, respectively [13]. While this extension is suitable for bottom-up sequencing with a single enzyme of narrow specificity, it should be applied with care otherwise, since an attempt to combine together sequence fragments of overlapping peptides will likely introduce inconsistencies in the offsets of the resulting aggregated strings.

Acknowledgements

The research by K.V. was partially supported by Government of Russian Federation (Grant 11.G34.31.0018, until December 2014) and Russian Science Foundation (Grant 14-50-00069, since February 2015). L.D. and M.V. are financially supported by The Netherlands Organization for Scientific Research (NWO), Zenith Grant 93511034. We are extremely grateful to Dr. Pavel Pevzner for fruitful discussions related to this our work, and for the anonymous referees for their fruitful comments and suggestions

Conflict of interest statement

The authors have declared no conflict of interest.

5 References

[1] Aebersold, R., Mann, M., Mass spectrometry-based proteomics. *Nature* 2003, 422, 198-207.

[2] Kelleher, N. L., Top down proteomics. *Anal. Chem.* 2004, 76(11), 197A-203A.

[3] Hu, Q., Li, H., Makarov, A., Hardman, M., Cooks, R. G., The Orbitrap: A new mass spectrometer, *J. Mass Spectrom.* 2005, 40(4), 430-433.

[4] Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A., Lajoie, G., PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* 2003, 17(20), 2337−42.

[5] Frank, A., Pevzner, P., PepNovo: De novo peptide sequencing via probabilistic network modeling. *Anal. Chem.* 2005, 77(4), 964−973.

[6] Chi, H., Sun, R. X., Yang, B., Song, C. Q., Wang, L. H., Liu, C., Fu, Y., Yuan, Z. F., Wang, H. P., He, S. M., Dong, M. Q., pNovo: De novo peptide sequencing and identification using HCD spectra, *J. Proteome Res.* 2010, 9(5), 2713-2724.

[7] Taylor, J. A., Johnson, R. S., Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom*. 1997, 11(9), 1067−1075.

[8] Dancik, V., Addona, T. A., Clauser, K. R., Vath, J. E., Pevzner, P., A. De novo peptide sequencing via tandem mass spectrometry. *J. Comput. Biol.* 1999, 6(3-4), 327−342.

[9] Pan, C., Park, B., McDonald, W., Carey, P., Banfield, J., VerBerkmoes, N., Hettich, R., Samatova, N., A high-throughput de novo sequencing approach for shotgun proteomics using high-resolution tandem mass spectrometry. *BMC Bioinformatics* 2010, 11, 118.

[10] Ma, B., Novor: Real-Time Peptide de novo sequencing software. *J. Am. Soc. Mass Spectrom.* 2015, 26(11), 1885-1894.

[11] Frank, A. M., Savitski, M. M., Nielsen, M. L., Zubarev, R. A., Pevzner, P. A., De novo peptide sequencing and identification with precision mass spectrometry. *J. Proteome Res.* 2007, 6(1), 114-123.

[12] Vyatkina, K., Wu, S., Dekker, L. J. M., VanDuijn, M. M., Liu., X., Tolić, N., Dvorkin, M., Alexandrova, S., Luider, T. M., Paša-Tolić, L., Pevzner, P. A., De novo sequencing of peptides from top-down tandem mass spectra. *J. Proteome Res.* 2015, 14(11), 4450-4462.

[13] Vyatkina, K., Wu, S., Dekker, L. J. M., VanDuijn, M. M. , Liu., X., Tolić, N., Luider, T. M., Paša-Tolić, L., Pevzner, P. A., Top-down analysis of protein samples by de novo sequencing techniques. *Bioinformatics* 2016, 32(18), 2753-2759.

[15] Vyatkina, K., De novo sequencing of top-down tandem mass spectra: A next step towards retrieving a complete protein sequence. *Proteomes* 2017, 5(1).

[15] Liu, X., Inbar, Y., Dorrestein, P. C., Wynne, C., Edwards, N., Souda, P., Whitelegge, J. P., Bafna, V., Pevzner, P. A., Deconvolution and database search of complex tandem mass spectra of intact proteins: a combinatorial approach. *Mol. Cell. Proteomics* 2010, 9(12), 2772-2782.

[16] Savitski, M., Nielsen, M. L., Zubarev, R. A., 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.* 2005, 4(8), 1180-1188.

[17] Gentzel, M., Köcher, T., Ponnusamy, S., Wilm, M., Preprocessing of tandem mass spectrometric data to support automatic protein identification. *Proteomics* 2003, 3(8), 1597-1610.

[18] Savitski, M. M., Mathieson, T., Becher, I., Bantscheff, M., H-Score, a mass accuracy driven rescoring approach for improved peptide identification in modification rich samples. *J. Proteome Res.* 2010, 9(11), 5511-5516.

[19] Frese, C. K., Altelaar, A. F. M., Hennrich, M. L., Nolting, D., Zeller, M., Griep-Raming, J., Heck, A. J., Mohammed, S., Improved peptide identification by targeted fragmentation using CID, HCD and ETD on an LTQ-Orbitrap Velos, *J. Proteome Res.* 2011, 10(5), 2377-2388.

[20] Senko, M. W, Beu, S. C., McLafferty, F. W., Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *J. Am. Soc. Mass Spectrom.* 1995, 6(4), 229-223.

[21] Zhang, Z., Marshall, A. G., A universal algorithm for fast and automated charge state deconvolution of electrospray mass-to-charge ratio spectra. *J. Am. Soc. Mass Spectrom.* 1998, 9(3), 225-223.

[22] Horn, D. M., Zubarev, R. A., McLafferty, F. W., Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *J. Am. Soc. Mass Spectrom.* 2000, 11(4), 330-332.

[23] Yuan, Z., Shi, J., Lin, W., Chen, B., Wu, F.-X., Features-based deisotoping method for tandem mass spectra. *Adv. Bioinformatics* 2011, Vol. 2011, Article ID 210805.

[24] Mann, M., Wilm, M., Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* 1994, 66(24), 4390-4399.

[25] Taylor, J. A., Johnson, R. S., Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry. *Anal. Chem.* 2001, 73(11), 2594−2604.

[26] Tabb, D. L., Saraf, A., Yates, J. R., GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. *Anal. Chem.* 2003, 75(23), 6415-6421.

[27] Sunyaev, S., Liska, A. J., Golod, A., Shevchenko, A., Shevchenko, A., MultiTag: multiple error-tolerant sequence tag search for the sequence-similarity identification of proteins by mass spectrometry. *Anal. Chem.* 2003, 75(6), 1307-1315.

[28] Searle, B. C., Dasari, S., Turner, M., Reddy, A. P., High-throughput identification of proteins and unanticipated sequence modifications using a mass-based alignment algorithm for MS/MS de novo sequencing results. *Anal. Chem.* 2004, 76(8), 2220-2230.

[29] Frank, A., Tanner, S., Bafna, V., Pevzner, P., Peptide sequence tags for fast database search in mass-spectrometry. *J. Proteome Res.* 2005, 4(4), 1287-1295.

[30] Tanner, S., Shu, H., Frank, A., Wang, L.-C., Zandi, E., Mumby, M., Pevzner, P. A., Bafna, V., InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Anal. Chem.* 2005, 77(14), 4626-4639.

[31] Cao, X., Nesvizhskii, A. I., Improved sequence tag generation method for peptide identification in tandem mass spectrometry. *J. Proteome Res.* 2008, 7(10), 4422-4434.

[32] Shen, Y., Tolić, N., Hixson, K. K., Purvine, S. O., Paša-Tolić, L., Qian, W.-J., Adkins, J. N., Moore, R. J., Smith, R. D., Proteome-wide identification of proteins and their modifications with decreased ambiguities and improved false discovery rates using unique sequence tags. *Anal. Chem*. 2008, 80(6), 1871-1882.

[33] Tabb, D. L., Ma, Z.-Q., Martin, D. B., Ham, A.-J. L., Chambers, M. C., DirecTag: Accurate sequence tags from peptide MS/MS through statistical scoring. *J. Proteome Res.* 2008, 7(9), 3838-3846.

[34] Perkins, D. N., Pappin, D. J. C., Creasy, D. M., Cottrell, J. S., Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 1999, 20(18), 3551-3567.

[35] Hoopmann, M.R., Finney, G.L., and MacCoss, M.J., High-speed data reduction, feature detection, and MS/MS spectrum quality assessment of shotgun proteomics data sets using high-resolution mass spectrometry. *Anal Chem* 2007, 79(15), 5620-5632.

[36] Hoopmann, M.R., MacCoss, M.J., and Moritz, R.L., Identification of peptide features in precursor spectra using Hardklör and Krönik. *Curr Protoc Bioinformatics* 2012, Chapter13:Unit13.18.

[37] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J., Basic local alignment search tool. *J. Mol. Biol.* 1990, 215(3), 403-410.

[38] Kim, S., Pevzner, P., MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.* 2014, 5, 5277.

[39] Olsen, J. V., Macek, B., Lange, O., Makarov, A., Horning, S., Mann, M., Higher-energy C-trap dissociation for peptide modification analysis. *Nat. methods* 2007, 4(9), 709-712.

Figure 1. Dependence of the number of peaks observed in isotope envelope on a peak mass for the bottom-up spectra deconvoluted with MS-Deconv. Masses are binned in 50 Da windows; for each bin and each envelope size observed for it, percent of the total number of peaks with a corresponding mass and size of the isotopic envelope is indicated. (Thus, the heights of all the columns sum up to 100%,)

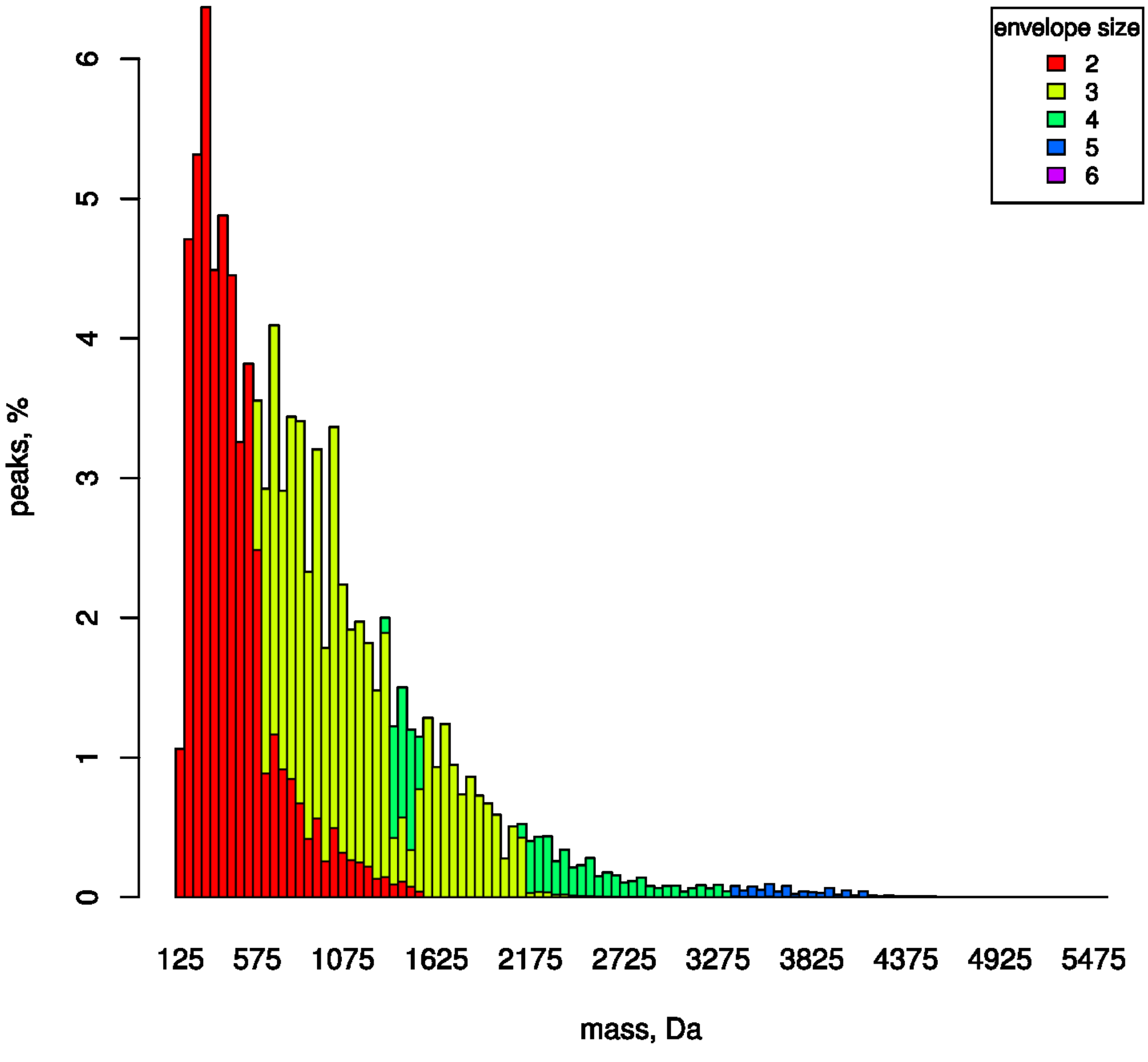
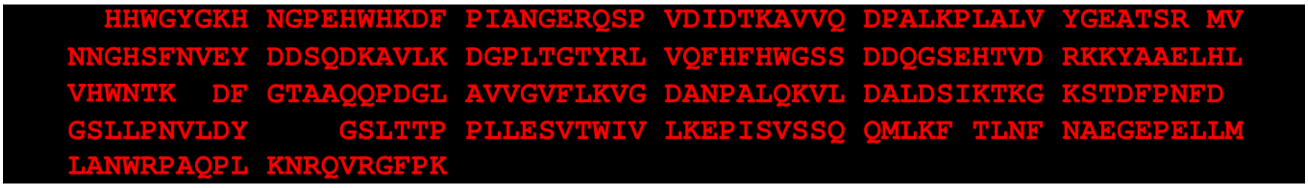


Figure 2. The coverage of the sequence of CAH2 with the correct fragments of length at least 4 of the *de novo* strings generated by Twister from the full dataset deconvoluted with MS-Deconv.



```
  HHWGYGKH NGPEHWHKDF PIANGERQSP VDIDTKAVVQ DPALKPLALV YGEATSR MV
NNGHSFNVEY DDSQDKAVLK DGPLTGTYRL VQFHFHWGSS DDQGSEHTVD RKKYAAELHL
VHWNTK  DF GTAAQQPDGL AVVGVFLKVG DANPALQKVL DALDSIKTKG KSTDFPNFD
GSLLPNVLDY     GSLTTP PLLESVTWIV LKEPISVSSQ QMLKF TLNF NAEGEPELLM
LANWRPAQPL KNRQVRGFPK
```

Table 1. Statistics on the number of peaks per spectrum for the full and tryptic datasets before and after deconvolution.

Dataset	Peaks per original spectrum			Deconvolution tool	Peaks per deconvoluted spectrum		
	Min	Max	Average		Min	Max	Average
Full	2	2137	328.24	MS-Deconv	0	115	17.57
				Xtract	0	522	45.54
Tryptic	2	1708	171.24	MS-Deconv	0	66	11.93
				Mascot Distiller	2	432	115.00
				Hardklör	1	464	31.83

Table 2. The coverage of the amino acid sequence of CAH2 with the entirely correct *de novo* strings generated from the full or tryptic dataset using different approaches.

Dataset	Method	Deconvolution tool	Details	*De novo* strings			Coverage for CAH2	
				Total	Correct		AAs	%
					#	%		
Full	Twister	MS-Deconv	4-tags, with peak reflection	14110	8720	61.80	238	91.54
		Xtract	4-tags, without peak reflection	12294	2537	20.64	230	88.46
	Vonode	-	Sequence Tags	139878	10209	7.30	173	66.54
		-	Consensus Sequence Tags	43153	4508	10.45	184	70.77
Tryptic	Twister	MS-Deconv	4-tags, with peak reflection	3164	1964	62.07	188	72.31
		Mascot Distiller	4-tags, without peak reflection	4315	886	20.53	196	75.38
		Hardklör	4-tags, with peak reflection	2483	733	29.52	204	78.46
	PepNovo	-	Top-scoring tag	12969	3627	27.97	169	65.00
	Novor	-		15937	169	1.06	70	26.92