Sequencing-Grade *De novo* Analysis of MS/MS Triplets (CID/HCD/ETD) From Overlapping Peptides

Adrian Guthals[1], Karl R. Clauser[3], Ari Frank[4], Nuno Bandeira[1,2,]*

[1]Department of Computer Science and Engineering, [2]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, California 92093; [3]Broad Institute of the Massachusetts Institute of Technology and Harvard, Cambridge, Massachusetts 02142; [4]Affectivon, Inc., Kiryat Tivon, Israel; [*]Corresponding author

Email: Adrian Guthals, aguthals@cs.ucsd.edu; Karl R. Clauser, clauser@broadinstitute.org; Ari Frank, ari.frank@gmail.com; Nuno Bandeira, bandeira@ucsd.edu

Key words: *de novo* sequencing, tandem mass spectrometry, peptide identification, protein sequencing

Full-length *de novo* sequencing of unknown proteins remains a challenging open problem. Traditional methods that sequence spectra individually are limited by short peptide length, incomplete peptide fragmentation, and ambiguous *de novo* interpretations. We address these issues by determining consensus sequences for assembled tandem mass (MS/MS) spectra from overlapping peptides (e.g., by using multiple enzymatic digests). We have combined Electron-Transfer Dissociation (ETD) with Collision-Induced Dissociation (CID) and Higher-energy Collision-induced Dissociation (HCD) fragmentation methods to boost interpretation of long, highly charged peptides and take advantage of corroborating b/y/c/z ions in CID/HCD/ETD. Using these strategies, we show that triplet CID/HCD/ETD MS/MS spectra from overlapping peptides yield *de novo* sequences of average length 70 AA and as long as 200 AA at up to 99% sequencing accuracy.

Introduction

In most proteomics studies, proteins are identified by digesting sample proteins into peptides (with an enzyme such as trypsin), generating a tandem mass (MS/MS) spectrum for each peptide precursor, and identifying the peptide sequence of each MS/MS spectrum with a database search tool, such as Sequest[1], Mascot[2], MS-GFDB[3], or Spectrum Mill[4]. Proteins IDs are then inferred from unique peptide sequence identifications. The utility of protein identification by database search depends upon the existence of a reference protein database that contains all proteins of interest. But due to mechanisms of sequence variation (such as genetic recombination and somatic hyper-mutation in monoclonal antibodies[5]) and the existence of unsequenced genomes, many protein sequences remain unknown. Nevertheless, the characterization of monoclonal antibodies and venoms from unsequenced species remains a key step in many therapeutic drug development pipelines[6–9]. Historically only a few low-throughput strategies have been available for *de novo* protein sequencing. As far back as 1987, Johnson and Biemann manually sequenced a complete protein from rabbit bone marrow using mass spectromtetry[10]. Edman degradation is another established approach for sequencing novel proteins but it has experimental bottlenecks which make it unsuitable for sequencing mixtures of proteins, proteins longer than 50 amino acids (AA), or posttranslationally modified proteins[11,12]. As such, many current applications of *de novo* sequencing still continue to rely upon manual curation of MS/MS spectra and/or Edman degradation[13–15].

Fully automated *de novo* strategies that interpret MS/MS spectra individually have been less successful compared to database search in part because they are limited by ambiguous interpretations of MS/MS fragmentation[16]. Even if both approaches use the same function for scoring peptide-spectrum matches (PSMs), the top scoring peptide in the database for a given MS/MS spectrum may be the 2nd or 7,000,000th highest scoring peptide over all possible *de novo* peptides, even if it is correct. Thus, *de novo* peptide sequencing algorithms typically report a ranked list of candidate PSMs for each spectrum where top-scoring PSMs have an accuracy of ~80-90% for low-resolution CID spectra[17,18] and ~90-92% for high-resolution CID spectra[16] (whereas database search results can typically be validated with 1% false discovery rate, FDR[19]). To yield these levels of accuracy, *de novo* tools face a significant tradeoff between sequencing accuracy and protein sequence coverage as spectra exhibiting complete peptide fragmentation rarely cover entire proteins, yet are required to reconstruct accurate sequences. *De novo* peptide sequencing approaches are also limited compared to low-throughput Edman methods in that they can only generate sequences as long as enzymatically digested peptides ($8-20$ AA) and thus cannot fully sequence protein(s) of interest.

An alternative approach to sequencing individual spectra is to simultaneously interpret multiple MS/MS spectra from overlapping peptides[20]. This Shotgun Protein Sequencing (SPS) paradigm has two distinct advantages over per-spectrum strategies. First, the alignment of spectra from overlapping peptides separates true N- and C-terminal ions from noise and leads to more accurate *de novo* sequences (~95% for high-resolution CID spectra) at almost full sequence coverage (95%)[21]. Second, the assembly of multiple aligned spectra allows for the extension of longer *de novo* sequences (up to 40 AA for high-resolution CID spectra)[21]. Remaining limitations of per-spectrum and SPS-based computational strategies have been addressed by incorporating imperfect databases of known proteins that are homologous to those in the sample. Depending upon the level of similarity between reference and target, an imperfect database can be used to correct *de novo* sequencing errors and anchor sequences to the reference (as done with Champs[22]), extend *de novo* sequences from known to unknown regions (as done with GenoMS[23]), or re-order *de novo* sequences to enable nearly full-length sequencing (as done with Comparative SPS, cSPS[24]).

De novo sequencing techniques have also been improved by utilizing multiple fragmentation modes. Compared to CID, alternative fragmentation strategies such as Higher-energy Collision Dissociation (HCD[25]) and Electron Transfer Dissociation (ETD[26]) are known to improve fragmentation and identification of long, highly-charged peptides[27]. HCD in particular has been shown to improve *de novo* peptide sequencing accuracy to ~95% and boost interpretations of long peptides, albeit at only 55% sequence coverage of peptides identified by database search[28]. When high-resolution CID and HCD spectra were processed with an updated SPS assembly algorithm (called Meta-SPS[29]), *de novo* protein sequences were extended to ~100 AA at the maximum and 20 AA on average at 94% sequencing accuracy/65% sequence coverage for a 6-protein sample mixture and 97% sequencing accuracy/89% sequence coverage for a purified monoclonal antibody. ETD has also been shown to improve per-spectrum sequencing length and accuracy[30], but the benefits of ETD for *de novo* sequencing are perhaps better utilized when it is paired with CID. In this approach, a

CID spectrum and an ETD spectrum are acquired for every precursor such that each pair of CID/ETD can be attributed to the same peptide. It is well known that CID and ETD exhibit complementary fragmentation patterns that, when paired with each other, can yield much richer N/C-terminal ion ladders for a greater variety of peptides[27]. Although the decreased scan rate of ETD means fewer MS/MS spectra can be acquired per aliquot of sample material, ETD significantly increases the fraction of identifiable spectra for both database search[3] and per-spectrum *de novo* sequencing[31,32], particularly when used in conjunction with enzymes such as LysC and GluC to acquire spectra from a greater variety of longer peptides (>20 AA)[33]. However, per-spectrum interpretation of paired fragmentation methods still cannot produce sequences longer than enzymatically digested peptides (13~20 AA depending on the digestion parameters) and has not achieved levels of sequencing accuracy/coverage greater than 95%/65% for high-resolution MS/MS[32]. Furthermore, published *de novo* sequencing tools capable of processing paired CID, HCD, or ETD spectra have not been made publicly available.

Advances in MS/MS instrumentation have enabled fast acquisition of a CID spectrum, HCD spectrum, and ETD spectrum per precursor such that each triplet of CID/HCD/ETD can be attributed to the same peptide. For example, a LTQ Velos Orbitrap instrument can acquire 5 triplets of CID/HCD/ETD MS/MS in a cycle of 1 MS in approximately the same amount of time as a cycle of 1 MS and 5 CID only MS/MS spectra on a prior generation LTQ-Orbitrap instrument. To take advantage of this capability, we describe a fully automated *de novo* protein sequencing approach that utilizes CID/HCD/ETD triplets from overlapping peptides to yield sequences as long as ~200 AA (~70 AA on average) at 99% sequencing accuracy and 71% sequencing coverage. To this end we updated algorithmic steps of the Meta-SPS[29] pipeline to process any combination of high-resolution CID, HCD, and ETD spectra from each peptide. Investigations into separate acquisition of CID, HCD, and ETD have showed promise for database search[34–36] but, to the best of our knowledge, this is the first application of triplet CID/HCD/ETD acquisition for de novo protein sequencing. We demonstrate that corroborating evidence of peptide fragmentation observed in CID/ETD pairs and CID/HCD/ETD triplets from overlapping peptides enables near-full length *de novo* protein sequencing at nearly perfect accuracy.

Procedures

Since Shotgun Protein Sequencing[21] interprets spectra from overlapping peptides, sample proteins were digested with multiple enzymes. High-resolution MS/MS CID/HCD/ETD triplets were then acquired on a Thermo LTQ-Orbitrap Velos and run through the updated Meta-SPS pipeline illustrated in Figure 1. To enable support for CID/HCD/ETD spectra we updated our pre-alignment steps to process and merge any combination of CID/HCD/ETD spectra from each precursor by adding two new stages to the Meta-SPS workflow. First PepNovo^{+}[16] was trained to score high resolution CID, HCD, and ETD MS/MS spectra (see section *PepNovo^{+} Training*). Since PepNovo^{+} cannot analyze multiple spectra from the same precursor, a procedure was developed to merge scored CID/HCD/ETD spectra and take advantage of corroborating evidence (see section *CID/HCD/ETD Merging*).

<Insert Figure 1>

MS/MS Acquisition

To benchmark and test this approach, 21,901 CID/HCD/ETD triplets (65,703 total MS/MS spectra) were separately acquired from aliquots of 7 digests of a mixture of 6 known proteins.

An equimolar mixture of 6 commercially purified proteins containing 252 ug of total protein was prepared. Cysteines were reduced with dithiothreitol (DTT) and alkylated with iodoacetamide. Seven 32 ug aliquots were created and used for 7 different digests with Trypsin, Chymotrypsin, Lys-C, Arg-C, Glu-C, Asp-N, or CNBr. The 6 proteins with accompanying molecular weights and Swiss-Prot accession numbers are bovine aprotinin (6.5 kDa, P00974), murine leptin (16 kDa, P41160), horse heart myoglobin (17 kDa, P68082), horseradish peroxidase (39 kDa, P00433), E. coli GroEL (57 kDa, P0A6F5), and human kallikrein-related peptidase (29 kDa, P07288). Details of sample preparation have been described previously[29].

Aliquots of each digest (~0.5 ug) were analyzed with an automated nano LC-MS/MS system, consisting of an Agilent 1200 nano-LC system (Agilent Technologies, Wilmington, DE) coupled to an LTQ-Orbitrap Velos Fourier transform mass spectrometer (Thermo Fisher Scientific, San Jose, CA) equipped with generation 2 ion optics (Velos Pro) and a nanoflow ionization source (James A. Hill Instrument Services, Arlington, MA). Peptides were eluted from a 10 cm column (Picofrit 75 um ID, New Objectives) packed in-house with ReproSil-Pur C18-AQ 3 um reversed phase resin (Dr. Maisch , Ammerbuch Germany) using a 95 min acetonitrile/0.1% formic acid gradient at a flow rate of 200 nl/min to yield ~20 sec peak widths. Solvent A was 0.1% formic acid and solvent B was 90% acetonitrile / 0.1% formic acid. The elution portion of the LC gradient was 3-6% solvent B in 1 min, 6-31% in 50 min, 31-60% in 13 min, 60-90% in 1 min and held at 90% solvent B for 5 min. Data-dependent LC-MS/MS spectra were acquired in ~3 sec cycles; each cycle was of the following form: one full Orbitrap MS scan at 60,000 resolution followed by 15 MS/MS scans in the orbitrap at 15,000 resolution using an isolation width of 3.0 m/z. The top 5 most abundant precursor ions were each sequentially subjected to CID, HCD, and ETD dissociation. Dynamic exclusion was enabled with a mass width of +/- 20 ppm, a repeat count of 1, and an exclusion duration of 12 sec. Charge state screening was enabled along with monoisotopic precursor selection and non-peptide monoisotopic recognition to prevent triggering of MS/MS on precursor ions with unassigned charge or a charge state of 1. For CID, the normalized collision energy was set to 30 with an activation Q of 0.25 and activation time of 30 ms. For HCD, the normalized collision energy was set to 45. For ETD, fluoranthene was used as the ETD reagent with an anion AGC target of 400,000 ions, supplemental activation was enabled, and the reaction time was dependent on the precursor charge state (precursor charge state - reaction time in msec: +2-100, +3-66.7, +4-50, +5-40, +6-33.3, etc). All MS/MS spectra were collected with an AGC target ion setting of 50,000 ions. The instrument control software does not currently allow for separate AGC targets for each dissociation mode. Optimal AGC targets would be closer to 30,000 ions for CID, HCD; and 200,000 ions for ETD[36]. All mass spectra associated with this paper may be downloaded from ftp://ftp.broadinstitute.org/distribution/proteomics/public_datasets/Guthals_JPR_2013.

Spectrum Preprocessing and Notation

Thermo RAW files were converted to mzXML with ProteoWizard[37] (version 3.0.3324). To validate *de novo* sequencing accuracy, all combinations of CID/HCD/ETD pairs/triplets as well as individual CID, HCD, and ETD spectra were searched with MS-GFDB[3] against the 6 target proteins and known contaminants with a spectrum-level false discovery rate of 1% (see Supplemental Materials for parameters used for MS-GFDB). As part of the Meta-SPS pipeline, high-resolution MS/MS peaks were first deconvoluted such that all peaks were converted to charge one[29]. The following notation is used below: a peptide MS/MS spectrum

S is defined as a collection of peaks where each peak $p \in S$ has mass $\mathrm{m}[p]$ and intensity $\mathrm{i}[p]$. The parent mass $\mathrm{M}[S]$ is the cumulative mass of all amino acids in the peptide sequence and the precursor charge $\mathrm{Z}[S]$ is the charge of the peptide precursor ion.

PepNovo$^+$ Training

Rather than processing MS/MS spectra directly, Meta-SPS uses PepNovo$^+$[16] to interpret MS/MS fragmentation patterns and convert MS/MS spectra into PRM (prefix residue mass) spectra where peak intensities are replaced with log-likelihood scores and peak masses are replaced by PRMs[38], or Prefix-Residue Masses (cumulative amino acid masses of N-term prefixes of the peptide sequence). Peak scores combine evidence supporting peptide *breaks* (observed cleavages along the peptide backbone, supported by either N- or C-terminal fragments). N/C-terminal fragments may be observed by b/y ions in CID/HCD and by $c/z°/z°$±H[39] ions in ETD. Because complementarity between b/y and $c/z°$ ions can cause C-terminal MS/MS ions to be misinterpreted as N-terminal ions, PRM spectra also typically contain many SRMs, or Suffix-Residue Masses (cumulative amino acid masses of C-terminal suffixes of the peptide sequence). This approach considers peaks in PRM spectra as both PRMs and SRMs because some spectra may contain predominantly SRMs and on average they make up 30-40% of all true PRMs or SRMs.

In previous work, high-resolution CID and HCD MS/MS spectra were scored with a PepNovo$^+$ scoring model that was not trained to process deconvoluted[29] spectra and there was no PepNovo$^+$ scoring model for ETD. In training the new models, we deconvoluted the training spectra because PepNovo$^+$ was optimized to analyze charge 2 and 3 tryptic CID spectra, and thus does not give enough weight to MS/MS peaks of charge 3 or higher in spectra from precursors of charge > 3. Here we trained three new scoring models[1] for deconvoluted high-resolution CID, HCD, and ETD MS/MS spectra using multiple data sets. The first consists of high-resolution CID, HCD, and ETD MS/MS spectra from tryptic peptides[36]. Another 175,595 tryptic HCD MS/MS spectra were provided by the Zubarev lab at the Karolinska Institute. The third data set consists of high-resolution ETD and HCD MS/MS spectra from Lys-C digestion and SCX fractionation of a yeast lysate collected in conjunction with the 2011 ABRF-iPRG study (see Supplemental Materials for description)[40]. All raw MS/MS spectra then were identified by MS-GFDB at 1% spectrum-level FDR to yield the set of training PSMs. PepNovo$^+$ used these PSMs to automatically learn ion types, intensity ranks, and noise models for each type of spectra and output models which can be used to score unidentified MS/MS spectra of the same type. See Supplemental Materials for details regarding the MS-GFDB searches and the specific PepNovo$^+$ training procedure.

CID/HCD/ETD Merging

Given a CID ($S^{\text{CID}} = \{c_1, \dots, c_n\}$), HCD ($S^{\text{HCD}} = \{h_1, \dots, h_m\}$), and/or ETD $\left(S^{\text{ETD}} = \{e_1, \dots, e_q\}\right)$ PRM spectra from the same precursor, the merging procedure generates a single merged PRM spectrum ($S = \{p_1, \dots, p_r\}$) (with the same parent mass $\text{M}[S]$) for all available spectra. Using the set of training PSMs, the objective is to maximize *observed breaks*, which is the percentage of all breaks observed as PRMs/SRMs at correct N/C-terminal masses (a measure of sensitivity), while also maximizing *explained score*, which is the percentage of

[1] These new models can only be used to generate PRM spectra, not *de novo* peptide sequences. Although PepNovo$^+$ PRM models were trained automatically with PSMs from >3,000 unique peptides per precursor charge state, training the rank-boosting[47] models needed for peptide sequencing required too many PSMs from unique peptides (>100,000) as well as more extensive modification of PepNovo$^+$ source code.

score in correct PRMs/SRMs relative to the score of all PRMs/SRMs in the same spectrum (a measure of accuracy). PRM spectra typically contain many C-terminal SRM masses along with N-terminal PRM masses. While PRM peaks have no offset from the summed amino acid masses, C-terminal peaks are offset by +18 Da (mass of H_2O) from SRMs in CID and HCD spectra[38,41]. In ETD spectra, C-terminal peaks are offset by -15 Da (mass of NH) from SRMs[3]. Given a PRM or SRM mass m, one can locate the complementary SRM or PRM mass in CID and HCD spectra with the formula $twin^{\mathrm{CID}}(m,S) = twin^{\mathrm{HCD}}(m,S) = \mathrm{M}[S] - m + 18$, while complementary masses in ETD can be found with $twin^{\mathrm{ETD}}(m,S) = \mathrm{M}[S] - m - 15$. Using these offsets, one can locate corroborating peaks from CID/ETD and HCD/ETD pairs that support the same peptide break, which are much more likely to explain true peptide breaks than individual PRMs. For example, we found that 92% of the score in peaks from identified ETD PRM spectra with matching peaks at the same (or complementary) mass in CID or HCD spectra was found in true PRMs/SRMs. In contrast, only 70-80% explained score is typically found in individual PRM spectra. Since PepNovo$^+$ does not currently recognize CID/HCD + ETD corroborating evidence when assigning log-likelihood scores, we post-processed the scores of corroborating PRMs/SRMs into combined scores in the merged PRM spectrum. However, since corroborating PRMs/SRMs only account for 47% of all peptide breaks in identified CID/HCD/ETD triplets, peaks without corroborating evidence must also be added to the merged spectrum.

Since 80% explained score was found to yield high *de novo* sequencing accuracy (97%) in a previous application of Meta-SPS[29], steps were developed to maximize the percentage of observed breaks at ≥ 80% explained score for all precursor charge states. First, corroborating PRMs and SRMs from CID/ETD and HCD/ETD pairs were extracted from PRM spectra and the corresponding combined PRMs were inserted into the merged spectrum. This was done in a series of steps to reduce the chances of misinterpreting SRMs as PRMs. But since steps [1-4] only captured PRMs and SRMs explaining 47% of all peptide breaks, the remaining peaks from CID, HCD, and ETD were also added to the merged spectrum in step [5] to bring the percentage of observed breaks to 94%. While this improved sensitivity, it also combined the noise between all three spectra such that the percentage of explained score was only 59% (instead of 91% for PRMs with corroborating evidence). Thus, local rank-based filtering was applied in step [6] to yield 86% observed breaks at 80% explained score over all precursor charge states (Figure 2b). We describe this procedure for merging CID/ETD pairs, but the same method can also be applied to HCD/ETD pairs.

[1] Consider all PRM/PRM matches: Find all pairs of peaks with same mass $(c_i, e_k : \mathrm{m}[c_i] = \mathrm{m}[e_k])$ and add a peak s to the merged spectrum S with PRM mass $\mathrm{m}[s] = \mathrm{m}[e_k]$. Whenever a peak is added to the merged spectrum, it only defines a new mass if that mass does not already exist in the merged spectrum within peak tolerance (otherwise the new peak's score is just added to the existing peak). Also find any complementary SRMs from the set $\{c_x, e_z : \mathrm{m}[c_x] = twin^{\mathrm{CID}}(\mathrm{m}[s], S) \wedge \mathrm{m}[e_z] = twin^{\mathrm{ETD}}(\mathrm{m}[s], S)\}$. For all of these peaks that were found, assign s the merged score $\mathrm{i}[s] = 2 \times (i[c_i] + i[c_x] + i[e_k] + i[e_z])$ and remove c_i, c_x, h_j, h_y, e_k, and e_z from S^{CID} and S^{ETD}, respectively.

[2] Consider all SRM/SRM matches with at least one PRM: Find all pairs of SRM peaks with mass difference 15+18 $(c_x, e_z : \mathrm{m}[c_x] = \mathrm{m}[e_z] + 33)$ *and* where at least one PRM from the set $\{c_i, e_k : \mathrm{m}[e_k] = twin^{\mathrm{ETD}}(\mathrm{m}[e_z], S) \wedge \mathrm{m}[c_i] = \mathrm{m}[e_k]\}$ is found from any spectrum (CID or ETD) for these SRMs. Then add a peak s to the merged spectrum S with the PRM

mass $\mathrm{m}[s] = \mathrm{m}[e_k]$, remove all of these peaks from S^{CID} and S^{ETD}, and assign s the merged score by the same formula in stage 1.

[3] Consider all PRM/SRM and SRM/PRM pairs: Find all pairs of PRM/SRM peaks $\left(c_i \in S^{\mathrm{CID}}, e_z \in S^{\mathrm{ETD}} : \mathrm{m}[c_i] = twin^{\mathrm{ETD}}(m[e_z], S)\right)$ or SRM/PRM peaks $\left(c_x \in S^{\mathrm{CID}}, e_k \in S^{\mathrm{ETD}} : \mathrm{m}[c_x] = twin^{\mathrm{CID}}(m[e_k], S)\right)$. Add a peak s to the merged spectrum with the PRM mass ($\mathrm{m}[s] = \mathrm{m}[c_i]$ for PRM/SRM pairs or $\mathrm{m}[s] = m[e_k]$ for SRM/PRM pairs), remove all of its supporting peaks from S^{CID} and S^{ETD}, and assign s the merged score by the same formula in stage 1.

[4] Consider all SRM/SRM matches without PRMs: Find all pairs of SRM peaks with mass difference 15+18 $(c_x, e_z : \mathrm{m}[c_x] = \mathrm{m}[e_z] + 33)$. Then add a peak s to the merged spectrum with the PRM mass $\mathrm{m}[s] = twin^{\mathrm{ETD}}(m[e_z], S)$, remove all of its supporting peaks from S^{CID} and S^{ETD}, and assign s the merged score by the same formula in stage 1.

[5] Add left over peaks from S^{CID} and S^{ETD} to S without changing their scores.

[6] Filter out peaks with low scores in S: a peak is retained if and only if its score is ranked in the top three over all neighboring PRM scores within a ±56 Da mass range.

The MS/MS spectra were acquired under conditions yielding mass measurement errors of +/-10 ppm for fragment masses and 30 ppm for parent masses. But since PepNovo$^+$ incorporates the parent mass error when assigning PRM masses from C-terminal fragment masses, a fixed 0.04 Da tolerance was used. This corresponds to 400 ppm @ m/z 100, 40 ppm @ m/z 1000, and 10 ppm @ m/z 4000. Merged PRM spectra from the same peptide were then clustered by an approach similar to MSCluster[42] (see Supplemental Materials for description). 21,901 CID/HCD/ETD triplets were combined into 11,325 clusters, each containing one or more triplets. A cluster contains only triplets sharing the same parent mass $\mathrm{M}[S]$. Thus, triplets derived from the same peptide, but in different precursor charge states, were still merged. Replicate triplet spectra exist in the dataset for two major reasons. First, given the small number of proteins in the sample and the rapid acquisition rate of the mass spectrometer, the dynamic exclusion time for triggering repeat acquisition of a particular precursor m/z was set to ~1/2 the chromatographic peak width to maximize the chance of collecting MS/MS near each peptide's chromatographic apex. Second, some of the same peptides can be produced by digestion with two different enzymes. For example some tryptic peptides are also produced by Lys-C or Arg-C digestion. The clustered set of merged PRM spectra was then run through the Meta-SPS pipeline illustrated in Figure 1, which involves two stages of alignment/assembly. PRM spectra were first aligned and assembled into contigs (sets of spectra from overlapping peptides)[21], which were further connected to form meta-contigs (sets of overlapping contigs)[29]. Figure 3 illustrates a resulting *de novo* protein sequence extracted from the highest-scoring consensus interpretation of a meta-contig.

<Insert Figure 3>

Results

The performance of Meta-SPS on CID/HCD/ETD triplets was assessed in terms of *de novo* sequencing length, coverage, and accuracy. Coverage and length was determined via modification-tolerant alignment of *de novo* sequences to the reference protein sequences[24]. Sequencing accuracy was also computed as done previously[29]: MS-GFDB peptide-spectrum matches were transferred to PRM spectra and then meta-contigs. A *sequence call* (mass of one or more possibly modified amino acids) was labeled *correct* if its consecutive flanking peaks are annotated by a MS-GFDB peptide match in the same ion series in the same identified spectrum (i.e. both are annotated as PRMs or SRMs from MS-GFDB's peptide match). All non-correct sequence calls from identified spectra are labeled *incorrect*. Remaining sequence calls whose flanking peaks are not from identified spectra are labeled *un-annotated*. See Supplemental Materials for details regarding the MS-GFDB searches used to compute performance metrics in Figure 2 and

Table 1. Figure 2a shows MS/MS ion statistics over all identified CID/HCD/ETD triplets and Figure 2c shows the numbers of identified spectra and peptides for all combinations of CID/HCD/ETD. Table 1a details the *spectrum coverage* by MS-GFDB (percent of protein sequence covered by identified peptides) for different combinations of fragmentation methods and Table 1b details coverage of all six proteins.

<Insert Figure 2>

<Insert Table 1>

Since Meta-SPS sequencing errors are usually distributed towards the ends of sequences[29] we removed the first and last sequence calls from every *de novo* sequence before computing coverage and accuracy. Resulting meta-contigs were binned by κ, the minimum allowable number of combined SPS contigs per meta-contig, and results are reported for $\kappa \geq 1$, $\kappa \geq 2$, and $\kappa \geq 5$. $\kappa \geq 5$ yields the longest and most accurate subset of meta-contig sequences because each of these must be supported by at least 5 SPS contig sequences, whereas $\kappa \geq 1$ retains un-merged SPS contigs with meta-contigs of all sizes to yield the highest sequencing coverage. At $\kappa \geq 5$, 19 *de novo* sequences assembling CID/HCD/ETD triplets were returned by Meta-SPS, all of which matched to the reference (with at most two modifications per match) and covered 71% of all six proteins at average length 66 AA (Table 1a). At $\kappa \geq 1$ and $\kappa \geq 2$, minimal losses in sequencing accuracy were sustained (98%) to achieve sequencing coverage (80% and 84%, respectively) closer to the coverage of database search (88%) at 1% FDR. The longest sequence spanned 194 AA and is shown in Figure 4 along with the longest sequences covering each of the six proteins.

Although sequences from CID/ETD pairs only (i.e., no HCD) were not as long at the maximum (125 AA), they were still longer than 50 AA on average (at $\kappa \geq 5$) and covered 67-

81% of target proteins depending on κ (Table 1a). HCD/ETD pairs exhibited roughly the same sequence coverage and length as CID/ETD (65-82% coverage, 131 AA maximum length, and 49 AA average length). The highest sequencing accuracy was observed for CID/ETD pairs and CID/HCD/ETD triplets at 99.5% and 98.9%, respectively, while HCD/ETD pairs gave 96.5% accuracy.

The last column in Table 1a displays sequencing statistics from separate acquisition of CID and HCD (11,010 high-resolution CID and 14,040 HCD MS/MS spectra after clustering)[29]. Even with spectra from ~2x as many precursors, this data set did not produce nearly as long and accurate sequences as did ETD paired with CID and/or HCD, which is explained by the increase in interpretable MS/MS fragmentation of long, highly charged peptides provided by ETD as well as the gain in PRM scores given to corroborating peaks in CID/ETD and HCD/ETD (Figure 2). Corroborating evidence was a very significant feature of peptide fragmentation as 91.8% of PRM scores was found in true PRMs after stage 1-4 merging. As a result, the combinations of CID/ETD, HCD/ETD, and CID/HCD/ETD gave the highest quality PRM spectra from long peptides, which are especially useful for assembly because they enable the extension of *de novo* sequences into regions that might not contain overlapping coverage of shorter peptides with precursor charge 2/3 due to either over-digestion or incomplete enzyme digestion. The quality of PRM spectra from long peptides was also improved by training PepNovo$^+$ on high-resolution CID, HCD, and ETD MS/MS spectra (Figure 2b). Surprisingly, the combination of CID/HCD gave slightly worse sequencing accuracy (91.2%) than separate acquisition of CID and HCD (93.6%), which is likely a result of the CID+HCD data set containing spectra from ~2x as many precursors: CID+HCD utilized two separate runs while the CID/HCD was acquired from a single run which also included time-consuming acquisition of ETD spectra (not used for this comparison between CID+HCD vs CID/HCD). Since peaks were merged in CID/HCD without corroborating evidence from ETD to determine PRM/SRM assignments, *de novo* sequencing quality was not increased enough to outweigh the loss of spectra from the diminished number of precursors subjected to MS/MS. It is noteworthy that CID/HCD exhibited roughly the same sequencing coverage as CID+HCD (Table 1a) and parameters can be set to improve CID/HCD sequencing accuracy at the cost of reduced coverage (parameters were set to the same values for CID/HCD and CID+HCD).

<Insert Figure 4>

Of the 6 proteins analyzed in this work, leptin and GroEL were produced recombinantly in *E. coli* while kallikrein-related peptidase, aprotinin, myoglobin, and peroxidase were isolated from natural sources. As documented in UniProt, leptin, kallikrein-related peptidase, aprotinin, and peroxidase are each known to contain N-terminal signal peptides that target the proteins for secretion from their cells of origin. Aprotinin and peroxidase further contain propeptide sequences that are cleaved upon activation. While the signal and pro-peptides would be missing from the proteins we analyzed, in Table 1 and Figure 4 we have used the full length gene sequence when calculating coverage by the assembled MS/MS spectra. Leptin contains a signal peptide (amino acids 1-21), that is lacking in the recombinant material obtained from Sigma-Aldrich[43]. Kallikrein-related peptidase contains a signal peptide (amino acids 1-17), a propeptide (amino acids 17-24), and known N-linked glycosylation at amino acid 69. Aprotinin contains a signal peptide (amino acids 1-21), and propeptides (amino acids 22-35 and 94-100). Peroxidase contains a signal peptide (amino acids 1-30), a propeptide (amino acids 339-353), and known N-linked glycosylation sites at (amino acids 43, 87, 188, 216, 228, 244, 285, and 298). The sugar micro-heterogeneity at N-linked glycosylation sites will tend to render any individual proteolytically-generated peptide containing that amino acid much less concentrated in the digestion mixture, and if subjected to MS/MS much less likely to yield interpretable fragmentation. These modifications, along with incomplete peptide sampling by the instrument, likely explain why 12% of protein sequences were not covered by database search. Remaining losses of coverage from *de novo* sequencing can be attributed to lack of spectra from overlapping peptides with sufficient fragmentation.

Conclusions

Multi-spectrum acquisition of high resolution CID, HCD, and ETD coupled with the proposed improvements to Meta-SPS enable near full-length automated *de novo* sequencing of simple protein mixtures at 99% sequencing accuracy. To the best of our knowledge, these are the longest and most accurate *de novo* sequences ever reported by an automated approach. Although this approach still falls short of fully reconstructing a complete protein, the average sequence length was greater than 60 AA long and approached 200 AA at the maximum, which should potentially enable automated sequencing of small proteins such as venom toxins[21,44] and the variable CDR regions of monoclonal antibodies[23,24,45].

Related methods for *de novo* sequencing with complementary fragmentation methods do not consider spectra from overlapping peptides, which limits sequencing length (<10 AA on average), accuracy (< 95%), and coverage (<70%) [31,32]. Still, results could possibly improve from devising more robust probabilistic scoring functions for paired CID/ETD and HCD/ETD MS/MS spectra than described here. Possible ways to do this include the Bayesian networks approach in Spectrum Fusion[31] or extensions of the scoring functions used in popular *de novo* tools like PepNovo$^+$ and PEAKS.

Although our high-resolution MS/MS acquisition enabled ±10 ppm peak tolerance, a fixed 0.04 Da tolerance was used because PepNovo$^+$ and SPS do not yet support ppm tolerance. Allowing for ±0.04 Da mass errors is equivalent to the diminishing mass error tolerance of 400-10 ppm over the increasing mass range of 100-4000 m/z. Implementing ppm tolerance in the Meta-SPS pipeline might allow for reduction alignment thresholds in SPS and Meta-SPS, as the probability of random high scoring matches between spectra from non-overlapping peptides diminishes with tighter mass tolerance. It would also enable resolving ambiguous interpretations of near isobaric masses (K-Q = 0.03638, K-GA = 0.03638, F-Mox = 0.0330, VS-W = 0.02113, and W-DA = 0.1526), which is a common limitation of proteomics mass spectrometry. Other ambiguities, such I/L interpretations, cannot be resolved by mass alone but may be resolved by examination of amino acid-specific fragmentation patterns[46].

This approach is mainly limited by instrument peptide sampling bias as a result of hydrophobicity, ionizability, and locations of basic amino acids, which leads to incomplete MS/MS coverage. This can significantly affect the performance of assembly-based approaches where full peptide coverage is not usable without sufficient overlap between peptides. As a result, Meta-SPS is currently optimized for datasets where the experimental protocol is expected to yield a high fraction of spectra from overlapping peptides. While this is currently easiest for simple protein mixtures, we would expect that the same methods would apply to more complex samples as long as enough mass spectrometry runs are used to acquire spectra from overlapping peptides. In addition, analysis of more complex mixtures would benefit from faster MS/MS scan rates or analysis of multiple fractions to yield enough coverage with multiple overlapping peptide sequences. The slower scan rate of ETD ($\approx \frac{2}{3}$ the rate of HCD) may further limit coverage, but our results suggest that ETD coupled with CID and/or HCD yields much longer and more accurate *de novo* sequencing than CID or HCD alone (even when considering that more precursors are subjected to MS/MS when fewer dissociation methods are employed), and thus the gains in sequencing outweigh the losses in peptide sampling. We further anticipate improvements in the quality of ETD spectra collected in the CID/HCD/ETD triplet configuration upon revision of the instrument control software to allow for separate AGC targets for each dissociation mode. Currently, we set the ETD AGC target ~4-fold lower than optimal so as not to overly compromise CID and HCD performance.

Acknowledgements

This work was partially supported by the National Institutes of Health Grant 8 P41 GM103485-05 from the National Institute of General Medical Sciences. This work was also supported in part by the Broad Institute of MIT and Harvard, and by grants to Steven A. Carr from the US National Cancer Institute (U24CA160034, part of the Clinical Proteomics Tumor Analysis Consortium initiative) and the National Heart, Lung, and Blood Institute (HHSN268201000033C and R01HL096738). We would like to thank Namrata Udeshi for providing expertise in configuring data acquisition, and the 2011 ABRF-iPRG committee for providing additional data that was generated in conjunction with their 2011 study. We would also like to thank the Zubarev lab at the Karolinska Institute for providing the HCD dataset used for training PepNovo$^+$.

References

1. Eng JK, McCormack AL, Yates JR (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry* 5:976–989.

2. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 20:3551–67.

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

4. Agilent Technologies, *Santa Clara, CA*:http://spectrummill.mit.edu/.

5. Noia JM Di, Neuberger MS (2007) Molecular mechanisms of antibody somatic hypermutation. *Annual review of biochemistry* 76:1–22.

6. Maggon K (2007) Monoclonal antibody "gold rush". *Current Medicinal Chemistry* 14:1978–1987.

7. Haurum JS (2006) Recombinant polyclonal antibodies: the next generation of antibody therapeutics? *Drug Discovery Today* 11:655–660.

8. Lewis RJ, Garcia ML (2003) Therapeutic potential of venom peptides. *Nature Reviews Drug Discovery* 2:790–802.

9. Pimenta AM, Lima ME De (2005) Small peptides, big world: biotechnological potential in neglected bioactive peptides from arthropod venoms. *Journal of Peptide Science* 11:670–6.

10. Johnson RS, Biemann K (1987) The primary structure of thioredoxin from Chromatium vinosum determined by high-performance tandem mass spectrometry. *Biochemistry* 26:1209–1214.

11. Thoma RS, Smith JS, Sandoval W, Leone JW, Hunziker P, Hampton B, Linse KD, Denslow ND (2009) The ABRF Edman Sequencing Research Group 2008 Study: investigation into homopolymeric amino acid N-terminal sequence tags and their effects on automated Edman degradation. *Journal of Biomolecular Techniques* 20:216–25.

12. Xiang B, Walters J, Mawuenyega K, Simpson J, Sandoval W, Smith JS, Hunziker P (2010) Results of the PSRG 2010 Study: Edman and Mass Spectrometric Terminal Sequencing of a Monoclonal Antibody. *Journal of Biomolecular Techniques* 21:S18.

13. Calvete JJ, Ghezellou P, Paiva O, Matainaho T, Ghassempour A, Goudarzi H, Kraus F, Sanz L, Williams DJ (2012) Snake venomics of two poorly known Hydrophiinae : Comparative proteomics of the venoms of terrestrial Toxicocalamus longissimus and marine Hydrophis cyanocinctus. *Journal of Proteomics* 75:4091–4101.

14. Medzihradszky KF, Bohlen CJ (2012) Partial De Novo Sequencing and Unusual CID Fragmentation of a 7 kDa, Disulfide-Bridged Toxin. *Journal of the American Society for Mass Spectrometry* 23:923–34.

15. Huancahuire-Vega S, Ponce-Soto LA, Martins-de-Souza D, Marangoni S (2011) Biochemical and pharmacological characterization of PhTX-I a new myotoxic phospholipase A2 isolated from Porthidium hyoprora snake venom. *Comparative biochemistry and physiology. Toxicology & pharmacology* 154:108–19.

16. Frank AM, Savitski MM, Nielsen ML, Zubarev RA, Pevzner PA (2007) De novo peptide sequencing and identification with precision mass spectrometry. *Journal of Proteome Research* 6:114–123.

17. Frank A, Pevzner PA (2005) PepNovo: de novo peptide sequencing via probabilistic network modeling. *Analytical chemistry* 77:964–73.

18. Ma B, Zhang K, Hendrie C, Liang C, Li M, Doherty-Kirby A, Lajoie G (2003) PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry* 17:2337–42.

19. Nesvizhskii AI (2010) A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *Journal of Proteomics* 73:2092–2123.

20. Bandeira N, Tang H, Bafna V, Pevzner P (2004) Shotgun protein sequencing by tandem mass spectra assembly. *Analytical chemistry* 76:7221–33.

21. Bandeira N, Clauser KR, Pevzner PA (2007) Shotgun protein sequencing: assembly of peptide tandem mass spectra from mixtures of modified proteins. *Molecular & Cellular Proteomics* 6:1123–34.

22. Liu X, Han Y, Yuen D, Ma B (2009) Automated protein (re)sequencing with MS/MS and a homologous database yields almost full coverage and accuracy. *Bioinformatics* 25:2174–80.

23. Castellana NE, Pham V, Arnott D, Lill JR, Bafna V (2010) Template proteogenomics: sequencing whole proteins using an imperfect database. *Molecular & Cellular Proteomics* 9:1260–70.

24. Bandeira N, Pham V, Pevzner P, Arnott D, Lill JR (2008) Automated de novo protein sequencing of monoclonal antibodies. *Nature Biotechnology* 26:1336–1338.

25. Olsen J V, Macek B, Lange O, Makarov A, Horning S, Mann M (2007) Higher-energy C-trap dissociation for peptide modification analysis. *Nature Methods* 4:709–712.

26. Syka JEP, Coon JJ, Schroeder MJ, Shabanowitz J, Hunt DF (2004) Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America* 101:9528–9533.

27. Guthals A, Bandeira N (2012) Peptide identification by tandem mass spectrometry with alternate fragmentation modes. *Molecular & Cellular Proteomics* 11:550–7.

28. 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 (2010) pNovo: de novo peptide sequencing and identification using HCD spectra. *Journal of Proteome Research* 9:2713–2724.

29. Guthals A, Clauser KR, Bandeira N (2012) Shotgun protein sequencing with meta-contig assembly. *Molecular & Cellular Proteomics* 10:1084–96.

30. Liu X, Shan B, Xin L, Ma B (2010) Better score function for peptide identification with ETD MS/MS spectra. *BMC Bioinformatics* 11:S4.

31. Datta R, Bern M (2009) Spectrum Fusion: Using Multiple Mass Spectra for De Novo Peptide Sequencing. *Journal of Computational Biology* 16:1169–82.

32. Savitski MM, Nielsen ML, Kjeldsen F, Zubarev RA (2005) Proteomics-grade de novo sequencing approach. *Journal of Proteome Research* 4:2348–2354.

33. Swaney DL, Wenger CD, Coon JJ (2010) Value of using multiple proteases for large-scale mass spectrometry-based proteomics. *Journal of Proteome Research* 9:1323–1329.

34. Shen Y, Tolić N, Xie F, Zhao R, Purvine SO, Schepmoes AA, Moore RJ, Anderson GA, Smith RD (2011) Effectiveness of CID, HCD, and ETD with FT MS/MS for degradomic-peptidomic analysis: comparison of peptide identification methods. *Journal of Proteome Research* 10:3929–43.

35. Shen Y, Tolić N, Purvine SO, Smith RD (2012) Improving Collision Induced Dissociation (CID), High Energy Collision Dissociation (HCD), and Electron Transfer Dissociation (ETD) Fourier Transform MS / MS DegradomeÀPeptidome

Identifications Using High Accuracy Mass Information Descriptions of Dat. *Proteome* 11:668–77.

36. Frese CK, Altelaar AFM, Hennrich ML, Nolting D, Zeller M, Griep-Raming J, Heck AJR, Mohammed S (2011) Improved peptide identification by targeted fragmentation using CID, HCD and ETD on an LTQ-Orbitrap Velos. *Journal of Proteome Research* 10:2377–88.

37. Kessner D, Chambers M, Burke R, Agus D, Mallick P (2008) ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* 24:2534–6.

38. Dancík V, Addona T a, Clauser KR, Vath JE, Pevzner P a (1999) De novo peptide sequencing via tandem mass spectrometry. *Journal of Computational Biology* 6:327–42.

39. Savitski MM, Kjeldsen F, Nielsen ML, Zubarev RA (2007) Hydrogen rearrangement to and from radical z fragments in electron capture dissociation of peptides. *Journal of the American Society for Mass Spectrometry* 18:113–20.

40. Clauser KR, Askenazi M, Bandeira N, Chalkley RJ, Deutsch E, Lam H, McDonald WH, Neubert T, Rudnick P, Martens L (2011) Proteome Informatics Research Group 2011 study. iPRG 2011: A Study on the Identification of Electron Transfer Dissociation (ETD) Mass Spectra. Available from: http://www.abrf.org/index.cfm/group.show/ProteomicsInformaticsResearchGroup.53.html.

41. Taylor J a, Johnson RS (2001) Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry. *Analytical chemistry* 73:2594–604.

42. Frank AM, Bandeira N, Shen Z, Tanner S, Briggs SP, Smith RD, Pevzner PA (2008) Clustering millions of tandem mass spectra. *Journal of Proteome Research* 7:113–122.

43. Sigma-Aldrich (2013) http://www.sigmaaldrich.com/.

44. Bhatia S, Kil YJ, Ueberheide B, Chait BT, Tayo L, Cruz L, Lu B, Yates JR, Bern M (2012) Constrained de novo sequencing of conotoxins. *Journal of Proteome Research* 11:4191–200.

45. Castellana NE, McCutcheon K, Pham VC, Harden K, Nguyen A, Young J, Adams C, Schroeder K, Arnott D, Bafna V, Grogan JL, Lill JR (2011) Resurrection of a clinical antibody: template proteogenomic de novo proteomic sequencing and reverse engineering of an anti-lymphotoxin-α antibody. *Proteomics* 11:395–405.

46. Gupta K, Kumar M, Chandrashekara K, Krishnan KS, Balaram P (2011) Combined Electron Transfer Dissociation-Collision-Induced Dissociation Fragmentation in the

Mass Spectrometric Distinction of Leucine, Isoleucine, and Hydroxyproline Residues in Peptide Natural Products. *Journal of Proteome Research* 11:515–22.

47. Frank AM (2009) A ranking-based scoring function for peptide-spectrum matches. *Journal of Proteome Research* 8:2241–2252.

Tables

a) Sequencing results for different fragmentation modes

		Fragmentation Modes	CID/HCD/ETD	CID/ETD	HCD/ETD	CID/HCD	CID+HCD*
		Spectrum Coverage (%)	88.3	86.9	87.9	88.3	87.4
		Longest Sequence (AA)	194	125	131	74	91
SPS contigs per meta-contig (κ)	κ ≥ 5	Sequencing Coverage (%)	70.9	67.3	64.8	52.7	52.8
		# Meta-Contigs	19	22	24	30	33
		Average Seq. Length (AA)	65.9	52.1	48.5	29.8	26.2
		Sequencing Accuracy (%)	98.9	99.5	96.5	91.2	96.3
		Un-annotated Seq. Calls (%)	4.1	1.3	5.2	8.2	2.3
	κ ≥ 2	Sequencing Coverage (%)	79.5	75.6	76.2	73.5	69
		# Meta-Contigs	33	38	45	50	61
		Average Seq. Length (AA)	43.9	37.4	34.3	28.9	20
		Sequencing Accuracy (%)	98.3	97.8	97	90.1	93.6
		Un-annotated Seq. Calls (%)	4	2.2	6.6	8.7	3
	κ ≥ 1	Sequencing Coverage (%)	83.6	80.9	82.1	81.7	75.3
		# Meta-Contigs	142	139	141	172	267
		Average Seq. Length (AA)	28.3	23.8	22.7	20.3	13.3
		Sequencing Accuracy (%)	97.9	97.3	96.4	88.5	86.6
		Un-annotated Seq. Calls (%)	4.2	3.4	7.5	11.4	4.8

*The CID+HCD column indicates sequencing results reported in a previous study of unpaired high-resolution CID and HCD MS/MS spectra from a Thermo LTQ Orbitrap using the same sample material and enzymatic digestions[29]

b) Sequencing results for triplet CID/HCD/ETD fragmentation

		Protein	leptin	kallikrein	groEL	myoglobin	aprotinin	peroxidase
		Protein Length (AA)	167	261	548	154	100	353
		Spectrum Coverage (%)	94.6	90.4	99.8	99.4	61	68.8
		Longest Sequence (AA)	93	134	194	80	59	58
SPS contigs per meta-contig (κ)	κ ≥ 5	Sequencing Coverage (%)	86.2	79.3	80.5	84.4	59	39.9
		# Meta-Contigs	3	5	4	2	2	4
		Average Seq. Length (AA)	66	60.2	115.8	65	39.5	35.2
		Sequencing Accuracy (%)	100	98.3	99.5	99.2	89.8	100
		Un-annotated Seq. Calls (%)	4.1	7.3	6.5	2.3	0	9.3
	κ ≥ 2	Sequencing Coverage (%)	86.2	87.7	92.3	84.4	59	53.8
		# Meta-Contigs	6	7	9	2	3	8
		Average Seq. Length (AA)	37.5	44.5	66.9	65	29.7	25.5
		Sequencing Accuracy (%)	100	98.4	97.7	99.2	89.8	100
		Un-annotated Seq. Calls (%)	7.3	5.9	2.8	0	9.3	2.2
	κ ≥ 1	Sequencing Coverage (%)	86.2	87.7	92.5	92.2	64	67.4
		# Meta-Contigs	10	12	16	4	5	17
		Average Seq. Length (AA)	26	29.5	41.9	37.5	20.4	17.1
		Sequencing Accuracy (%)	100	98.5	97.7	99.3	80	100
		Un-annotated Seq. Calls (%)	7.3	5.6	2.8	0	14.1	2.2

Table 1. *De novo* sequencing length, coverage, and accuracy for alternative minimum meta-contig size (κ) cutoffs. **(a) Sequencing results for different fragmentation modes:** *Spectrum Coverage* is the percent of amino acids in all proteins covered by peptides identified by MS-GFDB at 1% FDR. *Sequencing Coverage* is the percent of amino acids in all proteins covered by at least one aligned *de novo* sequence. *Average Seq. Length* is the average number of amino acids covered by each aligned *de novo* sequence and *Longest Sequence* is the maximum number of amino acids covered by a single *de novo* sequence. *Sequencing Accuracy* is the percentage of all annotated sequence calls that were labeled correct. *Un-annotated Seq. Calls* is the percentage of sequence calls that were un-annotated. Each column indicates which combination of MS/MS spectra was used as input to Meta-SPS and database search. **(b) Sequencing results for triplet CID/HCD/ETD fragmentation:** The same metrics in (a) are shown for each protein in the CID/HCD/ETD dataset (cumulative results over all six proteins are shown in the first column of (a)).

Figures

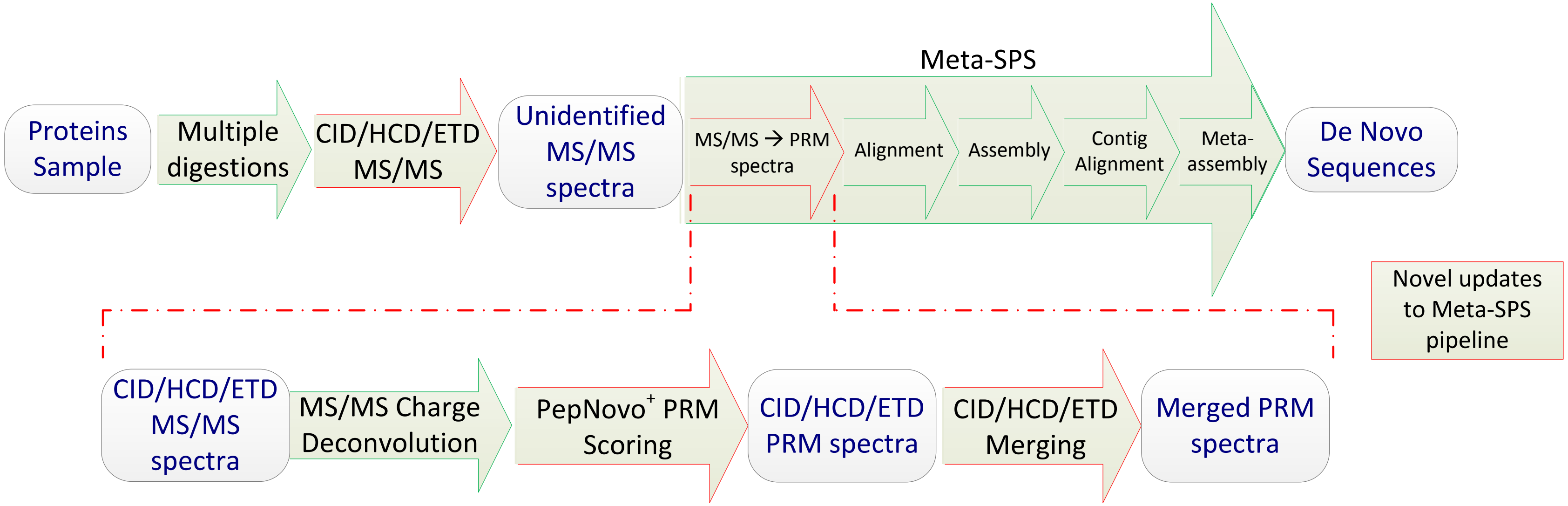
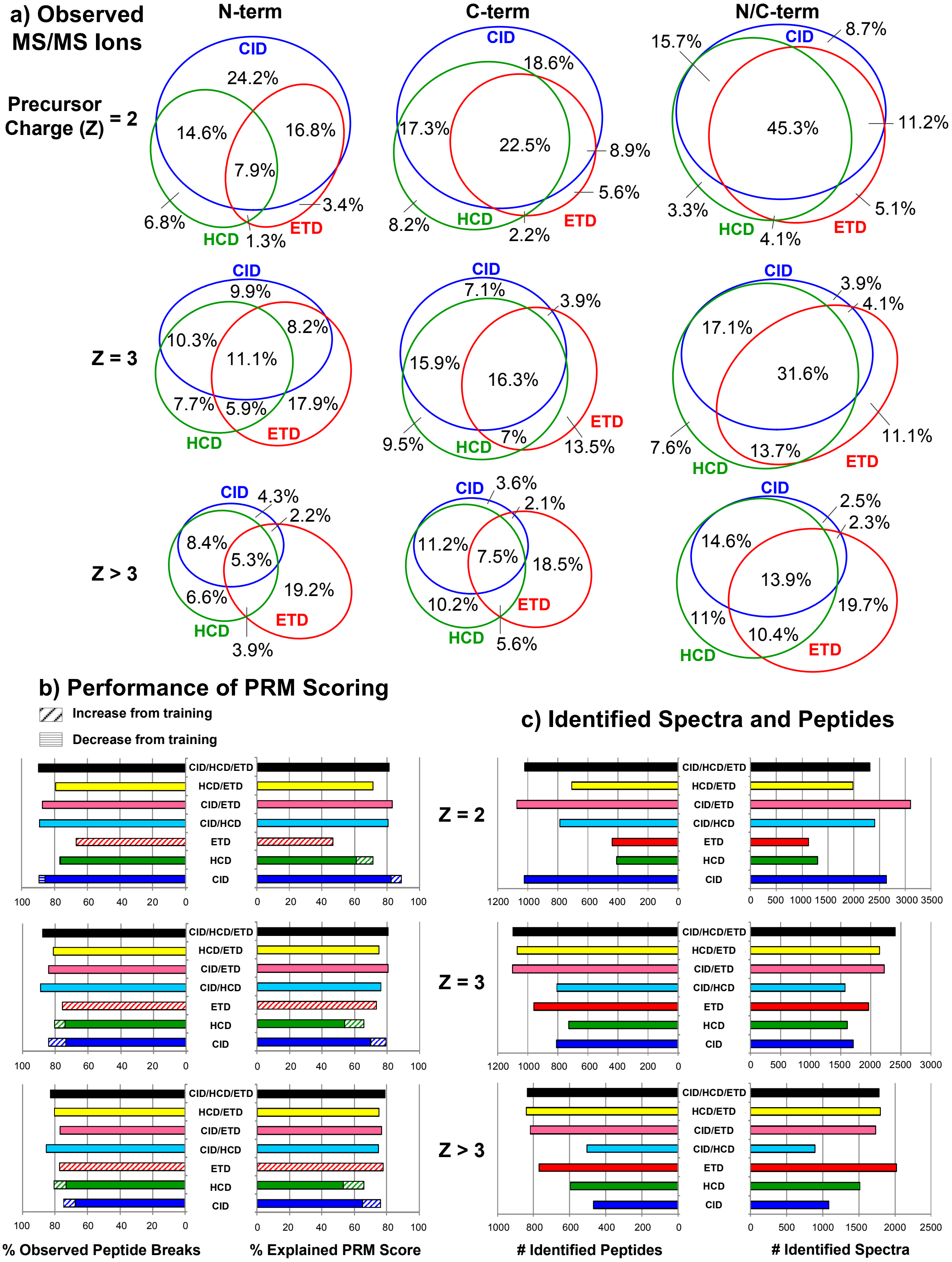


Figure 1. Updated Meta-SPS Pipeline: Green arrows denote procedures previously described in [29] and [21] while red arrows denote updated procedures described here.



a) Observed MS/MS Ions

b) Performance of PRM Scoring

c) Identified Spectra and Peptides

Figure 2. MS/MS ion statistics and performance of CID/HCD/ETD PRM scoring and merging. **(a) Observed MS/MS ions:** Percentage of peptide breaks observed by N-terminal ions (b ions in CID/HCD and c ions in ETD) and/or C-terminal ions (y ions in CID/HCD and $z°/z°$+H ions[39] in ETD) over all MS/MS CID/HCD/ETD triplets identified by MS-GFDB (considering a 10 ppm peak tolerance). To filter out low-intensity noise peaks, a peak was counted if and only if its intensity was ranked in the top seven over all neighboring peak intensities with in a ±56 Da radius. Rows separate baseline PSMs by precursor charge of identified triplets. **(b) Performance of PRM scoring:** Percentage of observed peptide breaks and percentage of explained score (the summed score of all true PRMs over the sum of all scores in the spectrum x 100) was counted over all combinations of merged/un-merged PRM spectra (without clustering) with identified MS/MS spectra. Peaks at N/C-terminal masses indicated peptide breaks in all cases. Each combination of PRM spectra was benchmarked by MS-GFDB IDs of the same combination of MS/MS spectra[3] (CID/HCD/ETD PRMs were benchmarked with CID/HCD/ETD IDs, CID/ETD PRMs with CID/ETD IDs, HCD PRMs with HCD IDs, etc). Also indicated is the performance gained by re-training PepNovo$^+$ to individually score high resolution CID, HCD, and ETD spectra. **(c) Identified spectra and peptides:** The numbers of identified spectra and unique peptides are shown for each combination of MS/MS spectra used to benchmark PRM scores in (b). As expected, incorporation of ETD significantly improves identification rates of spectra from highly charged precursors.

82 AA portion of myoglobin sequence

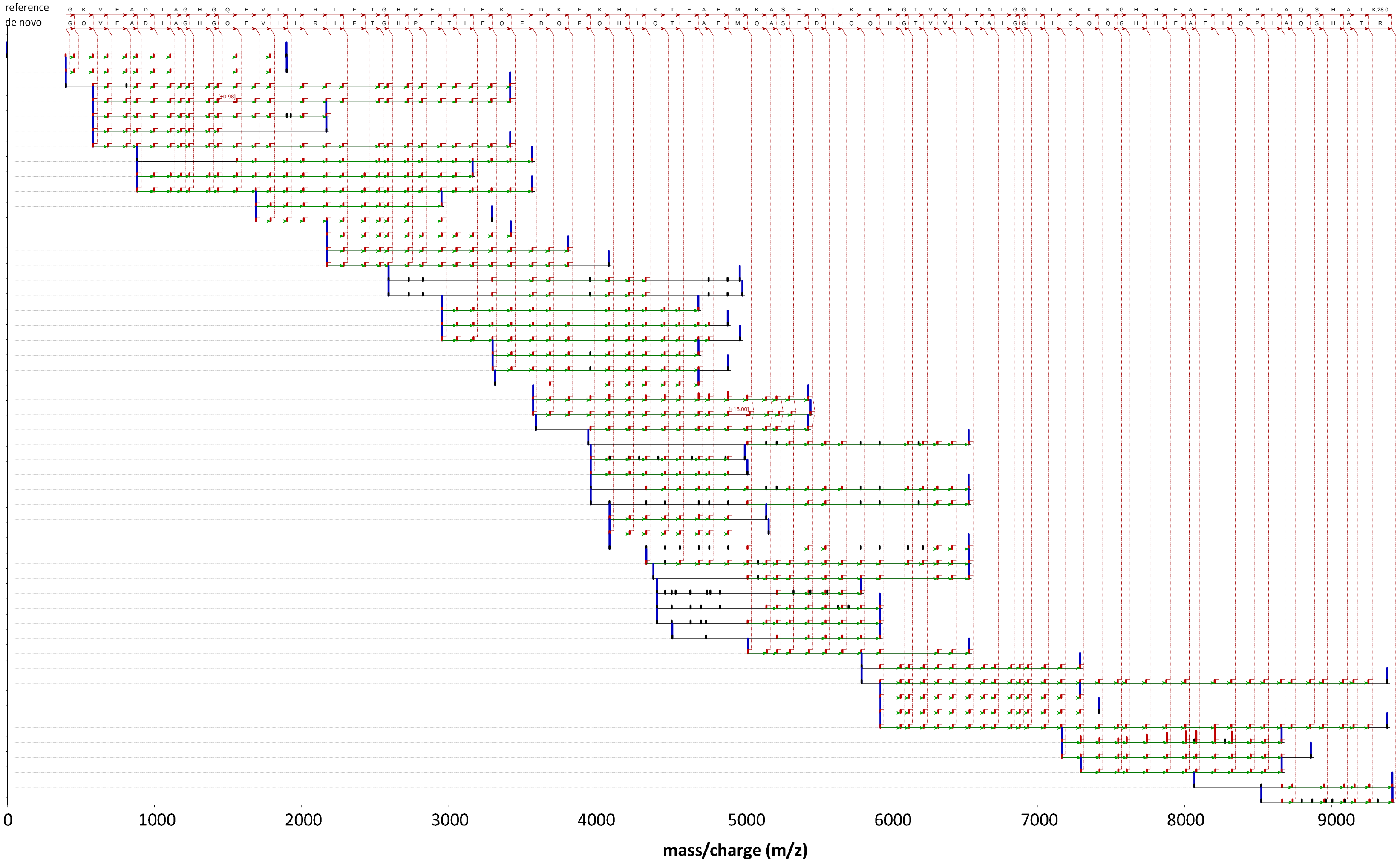


Figure 3. Assembled meta-contig of CID/HCD/ETD triplets: The top-most sequence is the myoglobin sequence as it is aligned to the *de novo* sequence below it. Each row denotes a merged PRM spectrum from one or more CID/HCD/ETD triplets where peaks not aligned to other merged PRM spectra from overlapping peptides are removed [21]. Red peaks indicate PRMs supporting the *de novo* sequence and green arrows between red peaks denote 1-2 AA mass differences supporting the consensus *de novo* sequence. Red vertical dotted lines connect assembled PRMs to each *de novo* sequence call; black peaks were not assembled into the consensus. Blue bars denote spectrum endpoints (at mass 0 and parent mass $\mathrm{M}[S]$). The height of each peak corresponds to the merged PRM score from CID, HCD, and ETD. The red labels "[+0.98]" and "[+16.00]" indicate post-translational modification masses that were tolerated during alignment/assembly (without knowing of them in advance). All *de novo* sequence calls, except the "R" at the end, were verified by database search.

Sequencing coverage of leptin

~~MCWRPLCRFLWLWSYLSYVQA~~VPIQKVQDDTKTLIKTIVTRINDISHTQSVSAKQRVTGLDFIPGLHPILSLSKMDQTLAVYQQV

VPIQKVQDDTKTLIKTIVTRINDISHTQSVSAXQRVTGLDFIPGLHPILSLSKM

SLSKMDQTLAVYQQV

LTSLPSQNVLQIANDLENLRDLLHLLAFSKSCSLPQTSGLQKPESLDGVLEASLYSTEVVALSRLQGSLQDILQQLDVSPEC

LTSLPSQNVLQIANDLENLRDLLHLLAFSKSCSLPQTSGLQKPESLDGVLEASLYSTEVVALSRLQGSLQDILQQLDVSP

S[KM]DQTLAVYQQVL[TS]LPSQNVLQIANDLENLRDLLHLLAFSK[SC]SL[PQ]TSGLQK[PE][SL]DGVLE[AS]LY[STE]VVALSRLQGSLQDILQQLDVSP

Sequencing coverage of kallikrein-related peptidase

~~MWVPVVFLTLSVTWIGAAPLILSR~~IVGGWECEKHSQPWQVLVASRGRAVCGGVLVHPQWVLTAAHCI~~RNK~~SVILLGRHSLFHPED

IVGGXECEKHSQPWQVLVASRGRAVCG

XXILLGRHSLFHPED

TGQVFQVSHSFPHPLYDMSLLKNRFLRPGDDSSHDLMLLRLSEPAELTDAVKVMDLPTQEPALGTTCYASGWGSIEPEEFLTPKK

TGQVFQVSHSFPHPLYDMSLLKNRFLRPGDDSSHDLMLLRLSEPAELTDAVKVMDLPTQEPALGTTCYASGWGSIEPEEFLTPKK

LQCVDLHVISNDVCAQVHPQKVTKFMLCAGRWTGGKSTCSGDSGGPLVCNGVLQGITSWGSEPCALPERPSLYTKVVHYRKWIKD

LQCVDLHVISNDVCAQVHPQKVTKFMLCAGRWTGGKS

XMLCAGRXTGGKSTCSGDSGGPLVCNGVLQGITSWGS

XSGDSGGPLVCNGVLQGITSWGSEPCALPERPSLYTKVVHYRKWIK

TIVANP

ILLGRHSLFHPEDTGQVFQVSHSFPHPLYDMSLLKNR[FL]RPGDDSSHDLMLLRLSEPAELTDAVKVMDL[PT]QE[PA]LGTTCYASGWGSIE[PE]E[FL]TPKKLQ[CV]D[LH]VISNDVCAQVH[PQ]KVTKFML[CA][GR][WT][GG]K

Sequencing coverage of groEL

~~M~~AAKDVKFGNDAHVKMLRSVNVLADAVKVTLGPKGRNVVLDKSFGAPTITKDGVSVAREIELEDKFENMGAQMVKEVASKANDAA

AAKDVKFGNDAXVKMLRXXNVLADAVKVTLGPKGRNVVLDKSFGAPTITKDGVSVAREIELEDKFENMGAQMVKEVASKANDAA

GDGTTTATVLAQAIITEGLKAVAAGMNPMDLKRGIDKAVTAAVEELKALSVPCSDSKAIAQVGTISANSDETVGKLIAEAMDKVG

GDGTTTATVLAQAIITEGLKAVAAGMNPMDLKRGIDKAVTAAVEELKALSVPCSDSKAIAQVGTISANSDETVGKLIAEAMDKVG

KEGVITVEDGTGLQDELDVVEGMQFDRGYLSPYFINKPETGAVELESPFILLADKKISNIREMLPVLEAVAKAGKPLLIIAEDVE

KEGVITVEDGTGLQDELDVVEGMQFDR

QFDRGYLSPYFINKPETGAVELESPFILLADKKISNIREMLPVLEAVAKAGKPLLIIAEDVE

GEALATLVVNTMRGIVKVAAVKAPGFGDRRKAMLQDIATLTGGTVISEEIGMELEKATLEDLGQAKRVVINKDTTTIIDGVGEEA

GEALATLVVXXXXXXXXKVAAVKAPGFGDRRKAM

XXLQDIATLTGGTVISEEIGMELEKATLEDLGQAKRVVINKDTTTIIDGVGEEA

AIQGRVAQIRQQIEEATSDYDREKLQERVAKLAGGVAVIKVGAATEVEMKEKKARVEDALHATRAAVEEGVVAGGGVALIRVASK

AIQGRVAQIRQQIEEATSDYDREKLQERVAKLAGGVAVIKVGAATEVEMKEKKARVEDALHATRAAVEEGVVAGGGVALIRVASK

LADLRGQNEDQNVGIKVALRAMEAPLRQIVLNCGEEPSVVANTVKGGDGNYGYNAATEEYGNMIDMGILDPTKVTRSALQYAASV

LADLRGQNEDQNVGIKVALR

AGLMITTECMVTDLPKNDAADLGAAGGMGGMGGMGGMM

AKDVKFGNDA(H,19)VKMLR(SV,-30)NVLADAVKVTLGPKGRNVVL[DK]SFGAPTITKDGVSVAREIELEDKFENMGAQMVKEVASKANDAAGDGTTTATV[LA]QAIITEGLKAVAAGMNPMDLKR[GI]DKAVTAAVEELKALSVPCSDSKAIAQVGTISANSDETVGKLIAEAMDKVGKEGVITVED[GTG]LQDELDVVEGMQFD

Sequencing coverage of myoglobin

~~M~~GLSDGEWQQVLNVWGKVEADIAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASEDLKKHGTVVLTALGGILKKKGHHEA

GKVEADIAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASEDLKKHGTVVLTALGGILKKKGHHEA

ELKPLAQSHATKHKIPIKYLEFISDAIIHVLHSKHPGDFGADAQGAMTKALELFRNDIAAKYKELGFQG

ELKPLAQSHATX

XXIPIKYLEFISDAIIHVLHSKHPGDFGADAQGAMTKALELFRNDIAAKYKE

KVEADIAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASEDLKKHGTVVLTALGGILKKKGHHEAELKPLAQSHAT

Sequencing coverage of aprotinin

~~MKMSRLCLSVALLVLLGTLAASTPGCDTSNQAKAQ~~RPDFCLEPPYTGPCKARIIRYFYNAKAGLCQTFVYGGCRAKRNNFKSAED

RPDFCLEPPYTGPCKARIIRYFYNAKAGLCQTFVYGGCRAKRNNFKSAED

CMRTCGGAIGPW~~ENL~~

CMRTCGGAIGPX

DFCL[EP]PYT[185.095][196.102]V[241.156]AQMYFYNAKAG[LC]QTFVYGGCRA[KR]NNFKSAEDCMRTCGGAIGP

Sequencing coverage of peroxidase

~~MHFSSSSTLFTCITLIPLVCLILHASLSDAQLTPTFYDNSCPNVSNIVRD~~TIVNELRSDPRIAASILRLHFHDCFVNGC~~DASILL~~

~~DN~~TTSFRTEKDAFGNANSARGFPVIDRMKAAVESACPRTVSCADLLTIAAQQSVTLAGGPSWRVPLGRRDSLQAFLDLANANLPA

RTEKDAFGNANSARGFPVIDRMKAAVESACPRTVSCADLLTIAAQQSVTLAGGPSWRVPL

GRRDSLQAFLDLANANLPA

PFFTLPQLKDSFR~~NVGLNR~~SSDLVALSGGHTFGKNQCRFIMD~~RLYNFSNTGLPDPTLNTTYLQTLRGLCPLNGNLSAL~~VDFDLRT

PFFTLPQLKDSFR

XXXXDLVALSGGHTFGKNQCRFX

PTIFDNKYYVNLEEQKGLIQSDQEL~~FSSPNATDTIPLVRSFANSTQTF~~FNAFVEAMDRMGNITPLTGTQGQIRLNCRVVNSNS~~LL~~

TFFNAFVEAMDRMGNITPLTGTQGQIRLNCRVVNSNS

~~HDMVEVVDFVSSM~~

TEKDAFGNANSARGFPVIDRMKAAVES[AC]PRTVSCAD[LL]TIAAQQSVTLAGG[PS]WRVP

Figure 4. *De novo* sequencing coverage of six target proteins at *κ* ≥ 5: Every colored row corresponds to a *de novo* sequence as separately mapped to the reference protein sequence (information not used by Meta-SPS); each row in the coverage map spans at most 85 AA. Regions of each sequence that were mapped to the reference with unknown modifications have X's in place of AA letter codes. Below each protein map is the longest *de novo* sequence covering that protein (also indicated in bold boxes in the coverage maps) following removal of first/last sequence calls. Blue letters correspond to calls that span 2 or more AA in the reference. Red letters indicate incorrect sequence calls as aligned to the reference. Remaining un-colored AA represent sequence calls that match reference amino acid masses. Regions where lack of *de novo* sequencing coverage was expected (due to lack of coverage by database search) are indicated with a dashed red line. As mentioned in the Results section, these lapses in coverage likely occur because of known cleavage of signal peptides and glycosylation sites.