Shotgun Protein Sequencing

ASSEMBLY OF PEPTIDE TANDEM MASS SPECTRA FROM MIXTURES OF MODIFIED PROTEINS*[S]

Nuno Bandeira‡§, Karl R. Clauser¶, and Pavel A. Pevzner‡

Despite significant advances in the identification of known proteins, the analysis of unknown proteins by MS/MS still remains a challenging open problem. Although Klaus Biemann recognized the potential of MS/MS for sequencing of unknown proteins in the 1980s, low throughput Edman degradation followed by cloning still remains the main method to sequence unknown proteins. The automated interpretation of MS/MS spectra has been limited by a focus on individual spectra and has not capitalized on the information contained in spectra of overlapping peptides. Indeed the powerful shotgun DNA sequencing strategies have not been extended to automated protein sequencing. We demonstrate, for the first time, the feasibility of automated shotgun protein sequencing of protein mixtures by utilizing MS/MS spectra of overlapping and possibly modified peptides generated via multiple proteases of different specificities. We validate this approach by generating highly accurate *de novo* reconstructions of multiple regions of various proteins in western diamondback rattlesnake venom. We further argue that shotgun protein sequencing has the potential to overcome the limitations of current protein sequencing approaches and thus catalyze the otherwise impractical applications of proteomics methodologies in studies of unknown proteins. ***Molecular & Cellular Proteomics 6:1123–1134, 2007.***

Current approaches to proteomics focus on the reliable identification of proteins under the assumption that all proteins of interest are known and present in a database. However, the limited availability of sequenced genomes and multiple mechanisms of protein variation often refute this assumption. Well known mechanisms of protein diversity include variable recombination and somatic hypermutation of immunoglobulin genes (1). The vital importance of some of these novel proteins is directly reflected in the success of monoclonal antibody drugs such as Rituxan™, Herceptin™, and Avastin™ (2, 3), all derived from proteins that are not directly inscribed in any genome. Similarly multiple commercial drugs have been developed from proteins obtained from species whose genomes are not known. In particular, peptides and proteins isolated from venom have provided essential clues for drug design (4, 5); examples include drugs for controlling blood coagulation (6–8) and drugs for breast (9, 10) and ovarian (11) cancer treatment. Even so, the genomes of the venomous snakes, scorpions, and snails are unlikely to become available anytime soon.

Despite this vital importance of novel proteins, the mainstream method for protein sequencing is still initiated by restrictive and low throughput Edman degradation (12, 13), a task made difficult by protein purification procedures, post-translational modifications, and blocked protein N termini. These problems gain additional relevance when one considers the unusually high level of variability and post-translational modifications in venom proteins (14, 15). Moreover the common labor-intensive approach of DNA cloning and sequencing from Edman chemistry-derived primers requires the additional availability of expensive instrumentation and expertise.

The primary function of venom is to immobilize prey, and prey animals vary in their susceptibility to venom. As a result, venom composition within snake species shows considerable geographical variation, an important consideration because snake bites (even by snakes of the same species) may require different treatments. Moreover the amount and number of different proteins and isoforms varies with gender, diet, etc. (16–18). These difficulties have been widely acknowledged (19, 20) and have motivated several attempts at *de novo* sequencing of MS/MS spectra from venom proteins (21, 22). However, all such attempts were made using traditional approaches that consider each MS/MS spectrum in isolation and thus face difficulties in the reliable interpretation of individual spectra (23–25).

Conceptually sequencing a protein from a set of MS/MS spectra can be described by a simple analogy. Imagine a jewelry box with many identical copies of a specific model of bead necklaces. Although all the beads are identical, this model is characterized by having irregular distances between consecutive beads; the set of interbead distances is initially chosen by the designer, and all necklaces are then made using exactly the same specification. Now assume that one day you open your jewelry box and realize that someone has vandalized all the necklaces by cutting them to fragments at randomly chosen bead positions. Can you recover the original design of this model of necklaces as specified by the set of consecutive interbead distances? In this allegory interbead distances correspond to amino acid masses, and beads cor-

From the ‡Department of Computer Science and Engineering, University of California San Diego, La Jolla, California 92093 and ¶Broad Institute of the Massachusetts Institute of Technology and Harvard, Cambridge, Massachusetts 02142

Received, January 2, 2007, and in revised form, April 4, 2007

Published, MCP Papers in Press, April 19, 2007, DOI 10.1074/mcp.M700001-MCP200

Downloaded from http://www.mcponline.org/ by guest on September 15, 2019

© 2007 by The American Society for Biochemistry and Molecular Biology, Inc.
This paper is available on line at http://www.mcponline.org

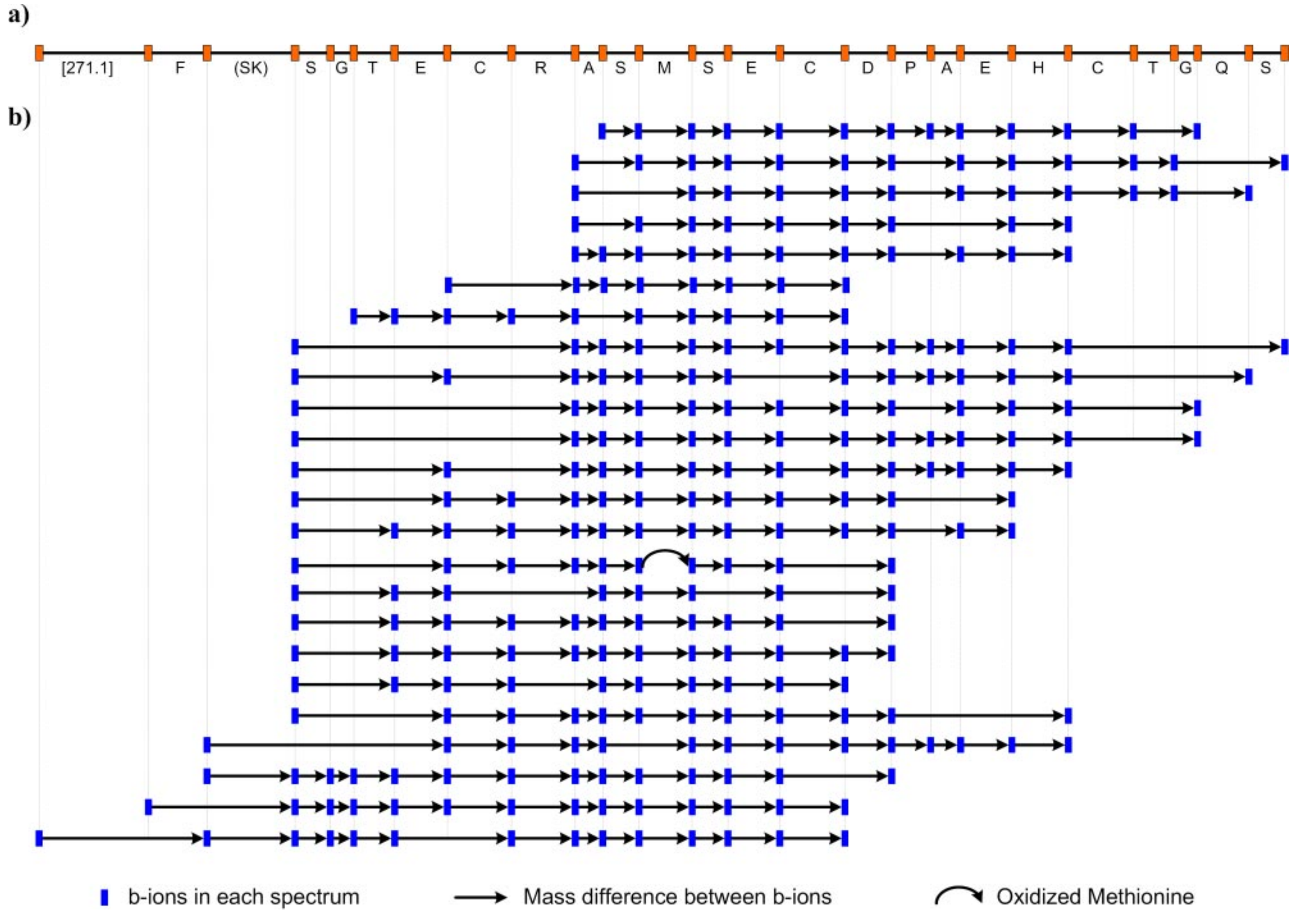


FIG. 1. **Contig assembling 24 spectra covering a 25-amino acid portion of *C. atrox* catrocollastatin.** Note that no single spectrum contains all the *b* ions for the recovered sequence even after we recovered missing *b* ions from correlated ion types (*e.g. y* ions). *a*, *de novo* contig sequence reconstructed from the assembled spectra. *b*, MS/MS spectra assembled in the contig. Each *line* corresponds to a different spectrum where matched *b* ions are shown as *blue rectangles* connected by *arrows*.

respond to MS/MS fragmentation points (between consecutive amino acids). MS/MS data add more than a few difficulties to this necklace assembly problem; for example, most peaks in MS/MS spectra do not correspond to any fragment ions (extra beads), and many fragment ions do not result in any peaks (missing beads). Nevertheless Fig. 1 presents an example of assembled MS/MS spectra resulting in an error-free 25-amino acid-long segment of catrocollastatin from western diamondback rattlesnake venom.

As far back as 1987, Klaus Biemann and co-worker (26) recognized the potential of tandem mass spectrometry for protein sequencing and manually sequenced a complete protein from rabbit bone marrow. In 2006, this approach was resurrected by Genentech researchers who were able to sequence antibodies by a combination of MS/MS and Edman degradation (27). With the same purpose in mind we have introduced previously a sequencing approach (28) that utilizes multiple MS/MS spectra from overlapping peptides generated using nonspecific proteases or multiple proteases with different specificities (29–31). Although this approach proved to be efficient for the assembly of a single purified unmodified protein, practical applications (like sequencing snake venoms) require applicability to mixtures of modified proteins. In fact, most MS/MS samples contain both modified and unmodified versions for many peptides, including both biological or chemical modifications introduced during sample preparation. However, it turned out that modifications present a formidable algorithmic challenge for assembly algorithms, and the performance of the approach in Ref. 28 degraded as soon as even a small percentage of the spectra come from modified peptides. To use the bead analogy, the necklace puzzle becomes very difficult if in addition to the canonical necklaces (non-modified proteins) the jewelry box also contains some necklaces that deviate from the designer's specification (modified proteins). In genomics, this challenge is not unlike that of assembling a highly polymorphic genome (like *Ciona* (32)), still an unsolved problem in bioinformatics.

Having recently developed an algorithm for the alignment of spectra from modified and unmodified peptide variants (33, 34), we now show that the integration of these alignments into shotgun protein sequencing is not trivial and indeed requires a completely new form of spectral assembly. To this end, we

Downloaded from http://www.mcponline.org/ by guest on September 15, 2019

introduce a generalized notion of *A-Bruijn graphs* (originally proposed in the context of DNA fragment assembly (35)) for the assembly of MS/MS spectra from overlapping modified and unmodified peptides into *contigs.* We further show how each contig then capitalizes on the corroborating evidence from the assembled spectra to yield a high quality *de novo* consensus sequence. In fact, comparison of our contig sequences with the protein sequences identified by standard database search reveals that shotgun protein sequencing results in the highest quality *de novo* interpretations ever reported for ion trap spectra from a mixture of modified proteins. Combined with an extensive contig coverage of the target proteins, our results indicate that the major remaining obstacle to high throughput protein sequencing is experimental rather than computational.

In genomics, DNA fragment assembly hardly ever produces a contiguous genome; even for small bacterial genomes it typically results in hundred(s) of disconnected contigs. Although these contigs cover almost the entire genomes, they are subject to *finishing* procedures that order and join contigs together using additional experiments. Similarly, limitations in proteolytic cleavage restrict shotgun protein sequencing to multiple contigs rather than contiguous proteins and motivate a quest for MS/MS-based (*e.g.* analysis of long multicharged peptides that connect different contigs) finishing experiments that would allow one to connect these contigs. Alternatively, exploratory results suggest that homology-tolerant comparison of contig sequences with known protein sequences may also be a viable approach for contig ordering (*i.e.* comparative protein sequencing).

Even in the absence of finishing experiments, our modification-tolerant approach readily generates much more information about western diamondback rattlesnake venom proteins than some of the most laborious Edman degradation/cloning studies (36). We obtained *de novo* sequences featuring 96% average coverage at an average sequencing accuracy of 90% and identified several polymorphisms and putative novel sequences with strong homology to known venom proteins from other snake species. We therefore argue that shotgun protein sequencing has the potential to overcome the limitations of current protein sequencing approaches and deliver a proteomics-based platform for studies of unknown proteins.

MATERIALS AND METHODS

The human inhibitor of nuclear factor κB kinase β (IKKβ)[1] dataset is a set of MS/MS spectra collected from multiple IKKβ samples and described in detail previously (37, 38). Briefly each sample was separately digested with different proteases (trypsin, elastase, and Glu-C) resulting in a rich ladder of spectra from overlapping peptides. IKK is known to be a key signaling complex involved in controlling cell proliferation, survival, and tumorigenesis (39). This IKKβ dataset was extensively analyzed with SEQUEST, Mascot, X!Tandem, and InsPecT (34, 38, 40) resulting in many reliably identified peptides and thus constitutes a gold standard against which to benchmark the performance of our sequencing approach. The IKKβ dataset contains 6126 reliably identified spectra from 524 unmodified peptides and 1383 reliably identified spectra from 346 modified peptides out of a total of 45,500 MS/MS spectra. We consider a spectrum to be reliably identified if it meets three criteria: (*a*) its InsPecT score is below the *p* value threshold for 5% false positives, (*b*) the spectrum contains at least 50% of all true *b* or *y* ions, and (*c*) at least 50% of the spectrum intensity is in *b*/*y* ions. The unusually high percentage of modified peptides (40% of all identified peptides were found to be modified) makes this a challenging dataset in our sequencing context. Beyond the usual artifactual modifications, this dataset additionally contains evidence (40) for Fe(III) adduct on Glu, sodium adducts on multiple residues including Gln, dehydration of Thr, a putative mutation of Ser to Asp, etc.

Venom Digestion and Mass Spectrometry—Our second dataset was generated from a sample of lyophilized *Crotalus atrox* western diamondback rattlesnake venom (Sigma-Aldrich). This venom was chosen for benchmarking our approach because it is relatively well studied, and several of its approximately two dozen proteins, ranging from 5 to 70 kDa, have been sequenced previously. The complexity of our sample is illustrated in an SDS-PAGE snapshot provided in our supplemental materials. Briefly the sample was reduced with DTT, and the cysteines were alkylated with iodoacetamide. The proteins that had not already precipitated were further precipitated with 60% ice-cold ethanol. After centrifugation, the supernatant was removed and discarded. The pellet was washed several times with 95% cold ethanol and then resuspended in 0.1% Rapigest (acid-labile SDS-like detergent). Four aliquots were created and diluted for 2-h digestions at pH 8.0 in 100 mM NH_4HCO_3; trypsin and Lys-C digests were performed in 0.085% Rapigest; chymotrypsin and Asp-N digests were performed in 0.01% Rapigest. Digestions were stopped, and the detergent was cleaved by acidifying with TFA, pH ~2. LC/MS/MS data were collected twice for each digest with an automated nano-LC/MS/MS system using an 1100 series autosampler and nanopump (Agilent Technologies, Wilmington, DE) coupled to either an LTQ or an LTQ-FT hybrid ion trap Fourier transform mass spectrometer (Thermo Electron, San Jose, CA) equipped with a nanoflow ionization source. Peptides were eluted from a 75-μm × 10-cm PicoFrit (New Objective, Woburn, MA) column packed with 5-μm × 200-Å Magic C-18AQ reversed-phase beads (Michrom Bioresources, Inc., Auburn, CA) using a 100-min acetonitrile, 0.1% formic acid gradient at a flow rate of 250 nl/min to yield 30-s peak widths.

Centroid mode data-dependent LC/MS/MS spectra were acquired in 3-s cycles; each cycle was of the following form: one full MS scan followed by eight MS/MS scans in the ion trap using normal scan rate on the most abundant precursor ions subject to dynamic exclusion for a period of 120 s after two repeats. For the LTQ dataset the acquisition software was LTQ version 1.0 SP1, the full ion trap MS survey scan was at the normal scan rate, and charge state screening was not used. For the LTQ-FT dataset the acquisition software was LTQ-FT version 1.0, the full FT MS survey scan was at 100,000 resolution with an automatic gain control target of 200,000 ions, and precursor ions of unassigned charge were excluded from triggering MS/MS. Spectrum Mill version 3.02b was used to extract all MS/MS spectra from each LC/MS/MS run including the spectral processing steps of merging replicates having a precursor mass within ±1.4 *m/z* and eluting within ±15 s, quality filtering to retain spectra with a sequence tag length >1, assigning precursor charges, and correcting ^{13}C precursor *m/z* misassignments. Precursor charges were assigned by Spectrum Mill for 62% of LTQ spectra using a combination of additional precursor charge states present in the MS spectra, *b*/*y* pairing in MS/MS

[1] The abbreviations used is: IKK, inhibitor of nuclear factor κB kinase.

Downloaded from http://www.mcponline.org/ by guest on September 15, 2019

spectra, and absence of peaks above the precursor mass in MS/MS spectra. This yielded 21,520 LTQ MS/MS spectra and 29,223 LTQ-FT MS/MS spectra. All LTQ-FT precursor charge assignments were done by the Thermo acquisition software using isotope spacing in the high resolution MS spectra. Lapses in precursor *m/z* assignment for LTQ-FT spectra by both the Thermo acquisition time software and postprocessing with Spectrum Mill for low abundance precursor ions are evident in Supplemental Table 1 by precursor masses given to only two decimal places rather than the usual four. Nearly all high confidence interpretations in Supplemental Table 1 for the four-decimal place precursor *m/z* assignments exhibit mass errors <10 ppm as expected. The two-decimal place LTQ-FT low abundance precursor *m/z* values have poorer mass accuracy and also exhibit some ^{13}C misassignments. Further peak detection and deisotoping for each spectrum was done independently in subsequent programs as needed.

Interpretation of Venom Spectra Using Database Search—A database of 5510 snake proteins was obtained from Swiss-Prot (August 3, 2006) by selecting all proteins from the taxa Serpentes, including 33 proteins and fragments from *C. atrox.* These *C. atrox* proteins were sequenced over the years in various laboratories using laborious Edman degradation as the first step. The obtained peptides were often used to design probes for further cloning and DNA sequencing. This database was extended with 19 protein sequences from common contaminants and proteases and 5529 "decoy" shuffled versions of all protein sequences. MS/MS spectra were searched against the database using InsPecT (38) with a peptide mass tolerance of 2.5 Da, fragment peak tolerance of 0.5 Da and allowing for oxidation on methionine, deamidation on asparagine, pyro-Glu from N-terminal glutamine, and pyrocarbamidomethylcysteine from N-terminal cysteine (41). The decoy database was used to enforce a false discovery rate of 1%, and all retained peptides had an InsPecT-assigned *p* value of 0.01 or less. Proteins were identified by iteratively selecting the protein sequence that explained the most identified spectra (minimum of 10 spectra per protein); a complete list of identified peptides and proteins is given in our supplementary materials.

Pairwise Spectral Alignment—As usual in the analysis of MS/MS spectra, we used several preprocessing steps. In particular, we used parent mass correction, parent charge estimation, and clustering of multiple spectra from the same peptide as described previously (34). Furthermore we replaced every peak with its likelihood score (42). This scoring combines the intensity of each peak, *b/y* complementarity, and presence/absence of neutral losses into a single likelihood score. Also it has the additional effect of making every spectrum symmetric, a desirable transformation because we often cannot tell *ab initio* which peaks come from prefix fragments (*e.g. b* ions) and which come from suffix fragments (*e.g. y* ions).

In our necklace problem, one can only rely on matching interbead distances from overlapping fragments to reconstruct the original sequence of consecutive interbead distances. This matching is the exact purpose of the spectral alignment described here: to find pairs of spectra from overlapping peptides (spectral pairs). Conceptually this procedure is akin to aligning interbead distances in that we need to detect overlaps between MS/MS spectra without knowing the corresponding peptides.

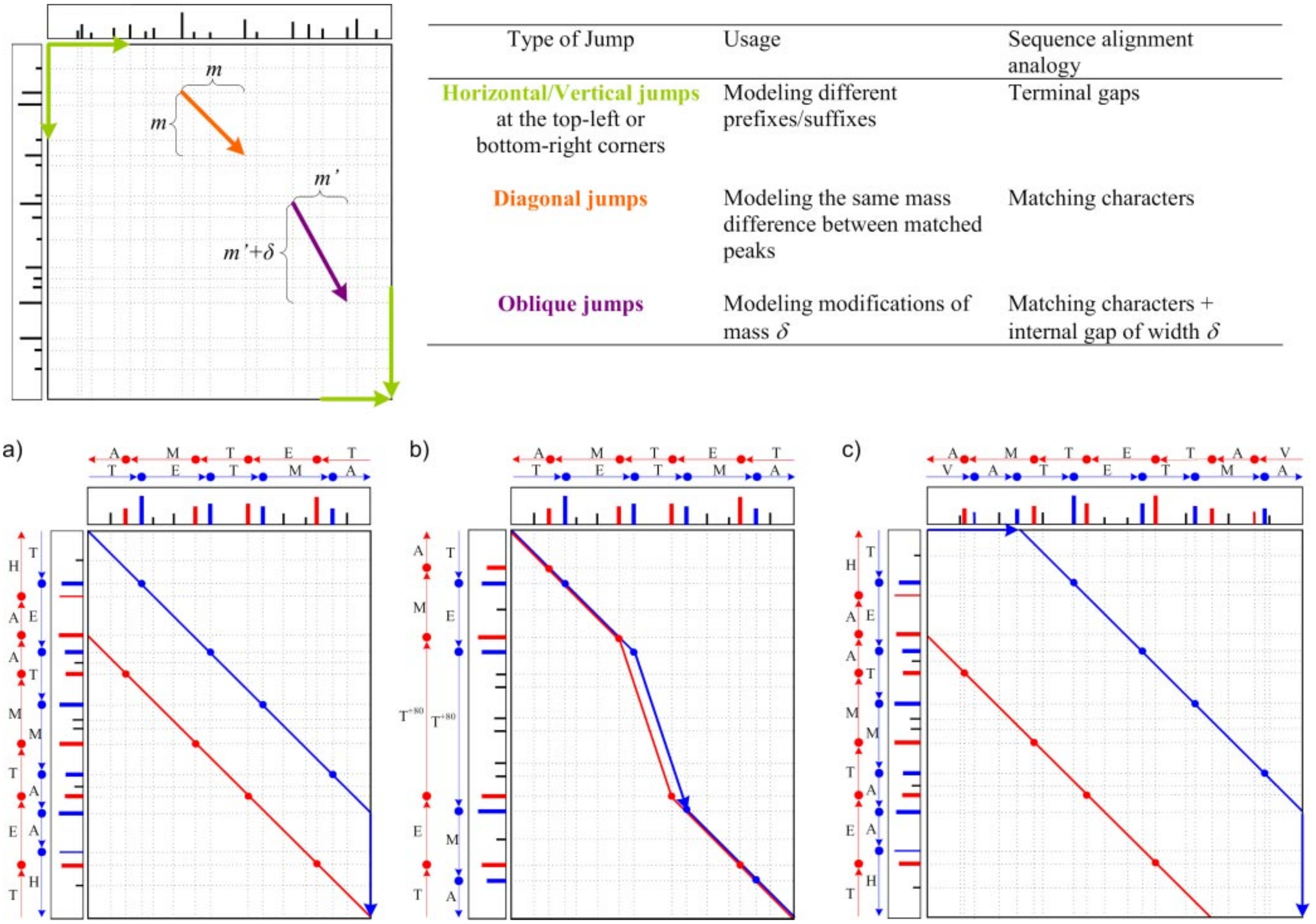
The algorithm for detection of spectra from overlapping peptides follows our previous approaches described previously (28, 33, 34, 40) (see Fig. 2). Spectral alignment translates the powerful Smith-Waterman sequence alignment technique (43) to the realm of MS/MS analysis. Like the dynamic programming matrix used in sequence alignment we construct a *spectral matrix* and find an *optimal path* in this matrix. Intuitively the spectral matrix of spectra S and S' is the set of pairs of peaks ($p \in S$, $p' \in S'$) called *matching peaks* (Fig. 2). Pairs of matching peaks may be connected by *jumps* as described in Fig. 2 with oblique jumps corresponding to putative modifications. As in classical sequence alignment, the optimal path (*i.e.* sequence of jumps) in the spectral matrix reveals the relationships between spectra. If spectra S and S' originate from overlapping peptides then there exists a path in this graph containing a large number of matching peaks, otherwise spectra S and S' are likely to be unrelated (in reality, peaks are scored by intensities as described previously (42)). Algorithmically spectral alignment is more complex than sequence alignment because in the former case one optimizes two correlated paths in the spectral matrix (one corresponding to *b* ions, illustrated in *blue*, and another corresponding to *y* ions, illustrated in *red*), whereas in the latter case one is only concerned with a single path. Although these paths are referred to as "blue" and "red" paths, in reality the colors of the paths are not known in advance. For clarity of exposition, we refer to blue diagonals in the text, whereas the algorithm works with "colorless" diagonals. These complications were recently addressed by our group (33, 34). We further note that although pairs of related spectra can also be identified by chemical tagging procedures (44, 45) or special instrumentation (46), these approaches do not consider overlapping peptides and cannot match spectra from multiple samples.

Fig. 2 presents three cases where spectral alignments help reveal overlapping and modified peptides from the IKKβ dataset without even trying to interpret the spectra: (*a*) SVSCILQEPK and SVSCILQEPKR (suffix extension), (*b*) SVSCILQEPK and SVSCILQ^{+22}EPK (modified variant), and (*c*) PESVSCILQEPK and SVSCILQEPKR (partial overlap). The corresponding optimal paths (shown in *blue* for *b* ions and *red* for *y* ions) and selected matching peaks between the different spectral pairs are illustrated in Fig. 2. Note that choosing where to place the jumps implicitly defines the type of spectral pair: modified/unmodified pair if there is an oblique jump in the middle, prefix/suffix pair if there is a single horizontal/vertical jump at the end/start, or overlap pair if there is one horizontal/vertical jump at the start and another at the end. The spectral alignment places the jump(s) in a position that maximizes the total scores of all matching peaks (34, 40). On a single desktop machine (Pentium4 at 2.8 GHz with 1 gigabyte of memory) our pairwise spectral alignment step executed in 46 min for the *C. atrox* dataset. However, the computation of pairwise spectral alignments can easily be executed in parallel and completed in only a few minutes when run on the University of California San Diego's FWGrid 64-node Linux cluster.

As a final step in our spectral alignment stage, we capitalize on a useful by-product of spectral alignment: the separation of *b* and *y* ions in the aligned spectra. Although the colors of the paths are unknown to the algorithm it turns out that, with high probability, the blue and red paths cleanly separate *b* and *y* ions. This separation is used to transform every aligned spectrum S into a *star spectrum*, a subset of S composed of mostly *b* ions or mostly *y* ions but not both. Star spectra were shown previously (34) to contain very few noise peaks while retaining most *b* ions (or *y* ions) and to be extremely selective of same-type ions (*i.e.* only *b* or only *y*).

Shotgun Protein Sequencing—It is widely accepted that *pairwise alignment whispers whereas multiple alignment shouts out loud*: combining pairwise spectral alignments into a single multiple alignment reveals peaks that are simultaneously supported by all or most of the aligned spectra. The high quality of star spectra may create the impression that the standard "overlap-layout-consensus" approach (47) for DNA fragment assembly should work for spectra assembly. In fact, we originally pursued this approach just to learn that it fails for MS/MS assembly as soon as even a small fraction of spectra represent modified peptides (28). The problem is that MS/MS spectra often come in *both* modified and unmodified versions thus posing a formidable challenge for assembly algorithms. In particular, the naïve overlap-layout-consensus approach simply projects all aligned peaks to a

Downloaded from http://www.mcponline.org/ by guest on September 15, 2019



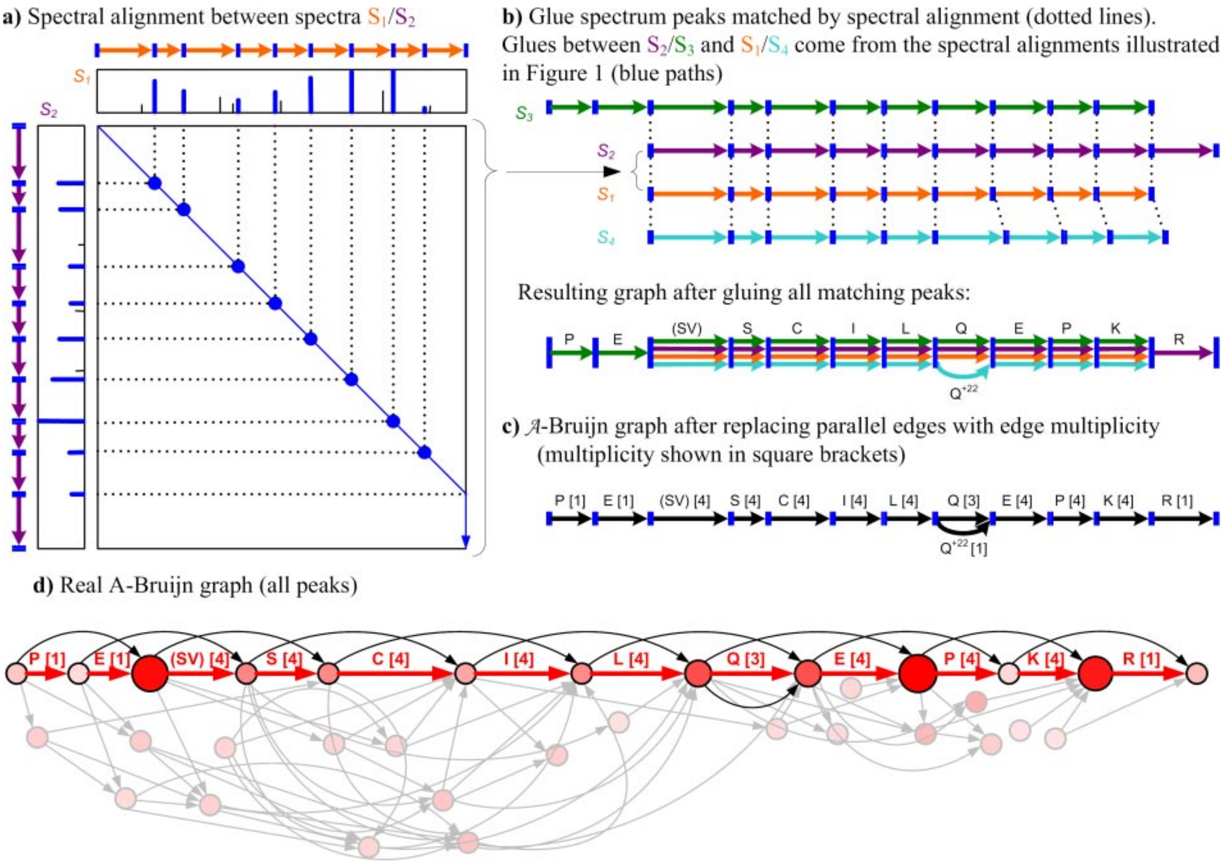
Type of Jump	Usage	Sequence alignment analogy
Horizontal/Vertical jumps at the top-left or bottom-right corners	Modeling different prefixes/suffixes	Terminal gaps
Diagonal jumps	Modeling the same mass difference between matched peaks	Matching characters
Oblique jumps	Modeling modifications of mass δ	Matching characters + internal gap of width δ

FIG. 2. **Pairwise spectral alignments (33, 34, 54) are computed with a dynamic programming algorithm similar to the Smith-Waterman sequence alignment algorithm (43); the corresponding intuitive interpretations are given in the table.** The alignment of two spectra is defined on the set of all matching peaks; each pair of matching peaks is represented as an *intersection* of *vertical* and *horizontal dotted lines* on the spectral matrix (*top left*). 18 peaks in the first spectrum and 17 peaks in the second spectrum result in 17 × 18 matching peaks in the spectral matrix. Matching peaks may be connected by three types of jumps: horizontal/vertical (because MS/MS spectra commonly lack peaks in the low/high mass regions, we also accept horizontal/vertical jumps to locations where no peaks are matched), diagonal, and oblique jumps. A spectral alignment is defined as a sequence of jumps from the *top left corner* to the *bottom right corner*. We consider spectral alignments with any number of diagonal jumps but a limited number of other jumps and distinguish between three types of spectral alignments: *a*, prefix/suffix alignments use a single horizontal/vertical jump (either at the *top left* or *bottom right*); *b*, modified/unmodified alignments use a single oblique jump; and *c*, partial overlap alignments use one horizontal/vertical jump at the *top left corner* and another at the *bottom right corner*. The optimal alignment of two spectra is an alignment with the longest sequence of valid jumps on the spectral matrix (the implemented scoring function is described in the main text). The alignment of *b* ions is shown in *blue*, and that of *y* ions is shown in *red*.

consensus spectrum and scores each consensus peak according to its co-occurrence in all overlapped spectra. Unfortunately this approach does not work when a set of overlapping spectra contains modifications because a simple projection of peaks onto a consensus spectrum would generate "shadow" peaks for each modification state. This shadowing effect would become even more severe if the alignment happened to include spectra from peptides with multiple modifications.

Note that although a spectral alignment is able to identify the mass and location of a modification, it is not immediately obvious which spectrum comes from the modified peptide, *i.e.* whether the modification corresponds to a loss or gain of residue mass. The situation becomes even more complex in the case of multiple modifications on the same peptide. Similar reasons help explain why assembly of *de novo* interpretations from the aligned spectra would lead to limited success at best. Even when no modifications are present, accurate *de novo* sequencing of MS/MS spectra is a difficult problem, often resulting in several possible peptides that explain the spectrum almost equally well. Thus, although committing any spectrum to a particular peptide would ignore the multiple alignment, considering all possible combinations of all top peptide interpretations would quickly lead to a combinatorial explosion of assembly configurations. However, the set of all possible interpretations of any given spectrum can be represented in a very compact way by a *spectrum graph*: each peak in the spectrum defines a vertex, and two vertices are connected by an edge if their peak masses differ by one or two amino acid masses (48). Also each vertex is assigned a score equal to the intensity of the corresponding spectrum peak. In this representation, every possible peptide interpretation corresponds to a path from zero to the parent mass of the spectrum (because there is a one-to-one correspondence between spectrum peaks and spectrum graph vertices, these terms will be used interchangeably). Fig. 3*a* illustrates two

Downloaded from http://www.mcponline.org/ by guest on September 15, 2019



a) Spectral alignment between spectra S_1/S_2

b) Glue spectrum peaks matched by spectral alignment (dotted lines). Glues between S_2/S_3 and S_1/S_4 come from the spectral alignments illustrated in Figure 1 (blue paths)

Resulting graph after gluing all matching peaks:

c) 𝒜-Bruijn graph after replacing parallel edges with edge multiplicity (multiplicity shown in square brackets)

d) Real A-Bruijn graph (all peaks)

FIG. 3. **Construction of an *A*-Bruijn graph from MS/MS data.** Star spectra of peptides SVSCILQEPK (S_1), SVSCILQEPKR (S_2), PESVSCILQEPK (S_3), and SVSCILQ^{+22}EPK (S_4) are "glued" together into an *A*-Bruijn graph using gluing instructions provided by pairwise spectral alignments shown in Fig. 2. *a*, the spectral alignment of spectra S_1 and S_2 shown in Fig. 2*a* reveals matching peaks in these spectra (only the blue path is shown). The peaks corresponding to *b* ions are shown in *blue*; other peaks are shown in *black*. Simplified spectrum graphs are shown next to each spectrum as paths through *b* ions. *b*, matching peaks in spectral alignments shown in Fig. 2, *a*, *b*, and *c*, generate pairwise gluing instructions between every pair of aligned spectra. Thus, *dotted lines* are used to represent both matching peaks in *a* and gluing instructions in *b*. *c*, parallel edges are replaced by a single edge with weight proportional to its multiplicity. In reality, edge weights are determined from peak intensities. *d*, real *A*-Bruijn graph using all peaks in the aligned spectra. Vertex scores are represented as *vertex size* and *color intensity*; edges to noise peaks are shown in *gray*. The path found by shotgun protein sequencing is shown in *red* with *edge labels* for the identified amino acids (*numbers* in *square brackets* indicate edge multiplicity).

simplified spectrum graphs for the aligned spectra S_1/S_2, showing only the vertices for the true *b* ions (in *blue*) and edges for the correct peptide path (in *orange* for S_1 and *purple* for S_2).

In the terms of the bead necklace analogy, each of these peptide paths would correspond to a necklace fragment from one of the original necklaces. Thus, we propose to reconstruct the original sequence of beads by finding similar pairs of overlapping fragments and "gluing" the matching beads to form a long chain identical to the original necklace model. Fig. 3 illustrates how this intuitive notion can be applied in the realm of spectral assembly: use spectral alignment to find the set of matching peaks between S_1/S_2 (Fig. 2*a*) and use these matches to glue the corresponding spectrum graph vertices (Fig. 2*b*). When applied to the simplified spectrum graphs in Fig. 3*a*, this would result in a merged spectrum graph with a single peptide path spelling the consensus sequence of S_1 and S_2. These merged spectrum graphs will be referred to as *A*-Bruijn graphs.

A-Bruijn graphs were first proposed by Pevzner *et al.* (35) in the context of repeat analysis and DNA fragment assembly. The key idea in their approach is to represent every DNA read as a path through nucleotides and "glue" all paths (reads) using matching nucleotides as pairwise gluing instructions. However, although each DNA read defines a single path through its nucleotide sequence, any given spectrum will correspond to a spectrum graph encoding many possible paths through its peaks. In fact, if genomic sequences did not contain so many similar and long repetitive regions, they would be much easier to assemble than protein sequences from MS/MS spectra! In particular, MS/MS spectra are intrinsically more error-prone than DNA reads: although reads are 98% accurate, MS/MS spectra contain mostly noise peaks, and the best known *de novo* peptide sequencing algorithms are only 75% accurate (23).

The process of using matching peaks to glue spectrum graphs into a single *A*-Bruijn graph is illustrated in Fig. 3. Note that edges between glued vertices are also glued if originally labeled with the same amino acid. Formally an *A*-Bruijn graph is constructed as follows: given a spectral alignment $S(S, S')$ on two spectra S and S' and two corresponding spectrum graphs G and G', output a single *A*-Bruijn graph $\mathcal{G}$ having G and G' as subgraphs. The specific gluing procedure is defined by the following operations.

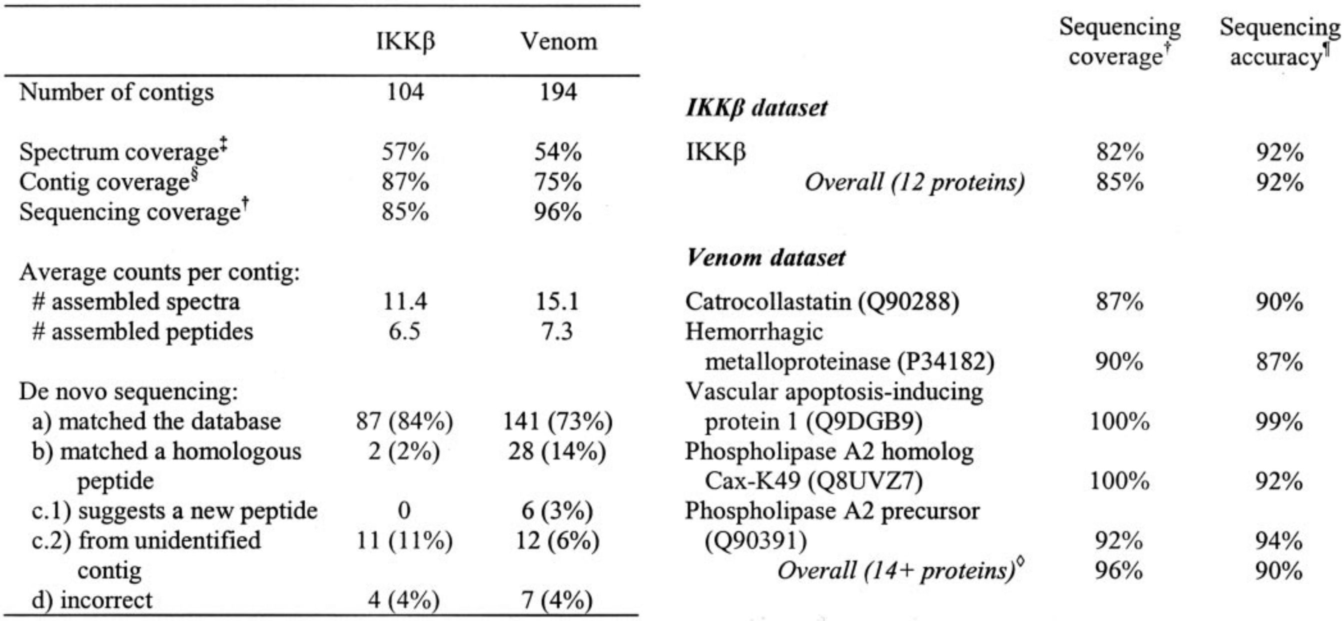
1. Vertices in $\mathcal{G}$: vertices $v_i \in G$ and $v'_j \in G'$ are glued into a single vertex in $\mathcal{G}$ if the corresponding peaks $p_i \in S$ and $p'_j \in S'$ are matched in $S(S, S')$. All remaining non-matched vertices are

Downloaded from http://www.mcponline.org/ by guest on September 15, 2019

TABLE I

Contigs obtained by shotgun protein sequencing

Types of contig sequences listed are: a, the contig sequence matched a protein that was expected to be in the sample; b, the contig sequence matched a peptide from an unexpected protein or suggested mutation of the target proteins; c.1, the contig sequence contains a tag of length ≥10 but did not match any peptide in UniProtKB, and the individual MS/MS spectra were not identified by database search (Spectrum Mill and InsPecT); c.2, like c.1 but containing only shorter tags; d, erroneous contigs (assembled spectra from non-overlapping peptides, or *de novo* sequence was incorrect).



	IKKβ	Venom
Number of contigs	104	194
Spectrum coverage‡	57%	54%
Contig coverage§	87%	75%
Sequencing coverage†	85%	96%
Average counts per contig:		
# assembled spectra	11.4	15.1
# assembled peptides	6.5	7.3
De novo sequencing:		
a) matched the database	87 (84%)	141 (73%)
b) matched a homologous peptide	2 (2%)	28 (14%)
c.1) suggests a new peptide	0	6 (3%)
c.2) from unidentified contig	11 (11%)	12 (6%)
d) incorrect	4 (4%)	7 (4%)

	Sequencing coverage†	Sequencing accuracy¶
IKKβ dataset		
IKKβ	82%	92%
Overall (12 proteins)	85%	92%
Venom dataset		
Catrocollastatin (Q90288)	87%	90%
Hemorrhagic metalloproteinase (P34182)	90%	87%
Vascular apoptosis-inducing protein 1 (Q9DGB9)	100%	99%
Phospholipase A2 homolog Cax-K49 (Q8UVZ7)	100%	92%
Phospholipase A2 precursor (Q90391)	92%	94%
Overall (14+ proteins)◊	96%	90%

‡ Spectrum coverage is the percentage of protein sequence represented in at least three spectra.

§ Contig coverage is defined as the assembled percentage of protein sequence represented in at least three spectra.

† Sequencing coverage is the percentage of contig regions that could be sequenced (in some instances there were not enough peaks in the assembled spectra to determine a complete amino acid sequence).

¶ Sequencing accuracy is defined as the percentage of correctly predicted amino acids.

◊ The venom dataset contained 14 reliably identified *C. atrox* proteins and provided strong evidence of containing additional, currently unknown venom proteins (described in the main text).

imported directly into $\mathcal{G}$. Each *A*-Bruijn vertex is scored by the sum of the intensities of the glued peaks.

2. Edges in $\mathcal{G}$: all edges in G and G' are imported directly into $\mathcal{G}$. However, edges are also glued if the end point vertices in G are glued to the end point vertices in G', and the edges are labeled with the same mass. Such pairs of edges, say e and e', are replaced by a single edge e'' of the same mass.

The construction of an *A*-Bruijn graph for a set of spectra and a set of spectral alignments is a straightforward iteration of the gluing operations described above. An example of a long sequence obtained from a set of 24 assembled spectra is illustrated in Fig. 1. However, errors in the spectral alignments may lead to the incorrect gluing of some peaks and generate inconsistent vertices in the *A*-Bruijn graph. In particular, it sometimes happens that multiple peaks from the same spectrum end up glued in the same vertex. Fortunately these inconsistencies are easily detected, and techniques are provided to resolve them (see supplemental materials).

After an *A*-Bruijn graph is constructed, the consensus sequence is defined as the heaviest path in the resulting directed graph. On most occasions, the resulting *A*-Bruijn graph is a directed acyclic graph, and thus standard algorithms are readily available to solve this problem. On the rare occasions when incorrect spectral alignments induce directed cycles in the *A*-Bruijn graph, we find that a simple greedy modification to the standard heaviest path algorithm works well on our *A*-Bruijn graphs (described in detail in our supplemental materials).

RESULTS

In the spirit of DNA fragment assembly (47), each set of overlapping spectra assembled by our approach is referred to as a *contig.* Table I lists the number of contigs assembled from each dataset along with some statistics on *A*-Bruijn graph construction and sequencing; *de novo* sequences obtained from the contigs are referred to as *contig sequences.* Note that contig sequences may be shorter than the span of amino acids covered by MS/MS spectra within a contig (some amino acids at the beginning/end of the contigs may not be recoverable). Overall these contig sequences covered 96% of all assembled regions in the venom dataset and 85% in the IKKβ dataset. Table I also shows the sequencing accuracy and coverage for the most abundant proteins in each dataset. It may appear that sequencing proteins is an easier task than sequencing DNA because protein sequences have few repeats or palindromes (the major source of difficulties in whole-genome assembly). However, not only are MS/MS spectra

Downloaded from http://www.mcponline.org/ by guest on September 15, 2019

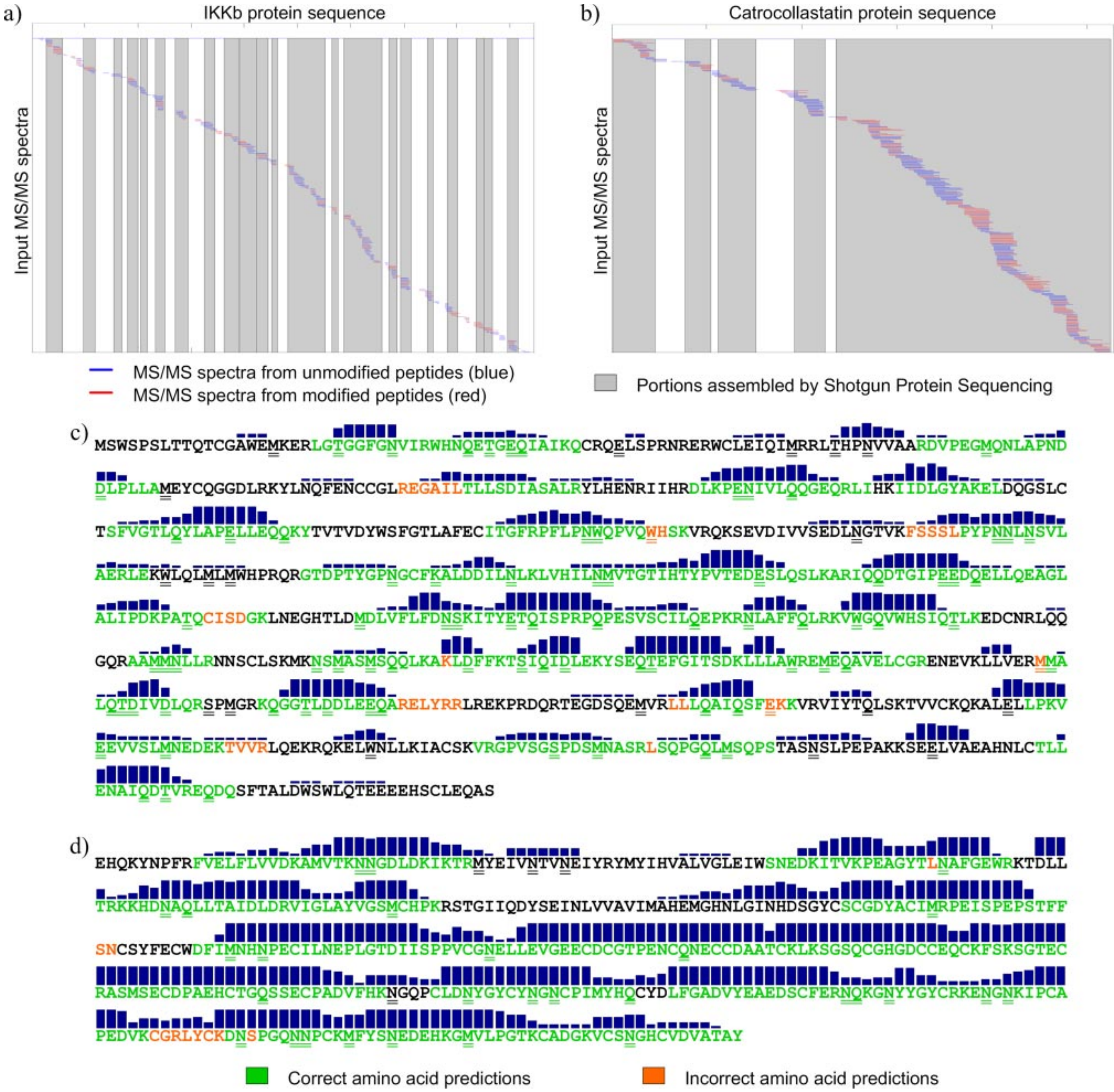


FIG. 4. Shown are assembled sets of spectra (*contigs*) for the most abundant protein in the IKKβ (*a*) and venom (*b*) samples. *Horizontal axes* represent amino acid positions, *vertical axes* represents the multiple spectra assigned to peptides from the corresponding protein, and each spectrum is represented as a *short blue/red horizontal line* for unmodified/modified peptides, respectively. *a*, 619 spectra from IKKβ resulted in 41 contigs. *b*, 1019 spectra from catrocollastatin precursor resulted in 34 contigs. *c*, recovered portions of the IKKβ protein sequence; correct portions are shown in *green* (430 amino acids), and incorrect portions are shown in *orange* (33 amino acids). The longest contiguous portion is 87 amino acids long, and 95% of its amino acids were correctly predicted. Amino acids found to be modified (oxidation, deamidation, dehydration, etc.) in at least one spectrum are shown *underlined*, and a *bar over* each amino acid indicates how often it occurred in the central portion (*i.e.* 20–80%) of all identified peptides; note that most *de novo* errors occur on non-central amino acids for which *b*/*y* peaks are often missing. *d*, recovered portions of the catrocollastatin protein sequence; correct portions are shown in *green* (321 amino acids), and incorrect portions are shown in *orange* (12 amino acids). The longest contiguous portion is 108 amino acids long, and all of its amino acids were correctly predicted. Note that the catrocollastatin protein in our sample is most likely a cleaved form of the sequence currently listed in Swiss-Prot (36).

intrinsically more error-prone than DNA reads, but peptide sampling is strongly biased and results in some portions of the proteins being represented in many spectra whereas others are not seen at all. As a result, the observed peptides often correspond to isolated sets of overlapping spectra separated by coverage gaps or sometimes connected by only one or two spectra. Fig. 4 shows the spectrum coverage observed for the IKKβ and catrocollastatin proteins (see supplemental materi-

Downloaded from http://www.mcponline.org/ by guest on September 15, 2019

TABLE II
Homologous contig sequences from the venom dataset

The segments identical to the *de novo* reconstructions are shown underlined. On the *de novo* sequences, parentheses indicate sequences where the order of the amino acids was not determined; square brackets indicate indistinguishable amino acid masses (on ion trap spectra). A homologous sequence is confirmed (√) if it matches the peptides obtained by independent traditional database search of the assembled MS/MS spectra. This confirmation step turned out positive whenever the homologous peptide was present in the database (albeit on a protein from a different snake species); assembled spectra in the remaining homologous contig sequences had no significant match to the database and were thus neither confirmed nor refuted. All *C. atrox* homologies were either matched to a different snake species or can be explained by single nucleotide polymorphisms of the original sequences, which were also detected in our sample. The complete list of all putative homolog peptides can be found in our supplemental materials as well as annotated MS/MS spectra for all novel homologies.

De novo sequence	Homologous matched	Homologous protein	Species, protein name
L(TP)GSQCAD(GV)CCDQCRF[Q,K]	LTPGSQCADGVCCDQCRFT	O42138√	*Agkistrodon contortrix laticinctus*, metalloproteinase-disintegrin-like protein
	LRPGSQCAEGMCCDQCRFM	Q2QA03	*C. durissus durissus*, metalloproteinase P-II
	LRPGAQCADGLCCDQCRFI	P68520	*C. atrox*, platelet aggregation activation inhibitor
KVLNEDEQTRD(PK)	KVLNEDEQTRDPK	Q9DF66√	*Trimeresurus jerdonii*, venom serine proteinase 3
	KVPNEDEQTRNPK	Q8QHK2	*C. atrox*, serine protease catroxase II
(LTNCSPK)(TD)IYSYSWKR	LTNCSPKTDIYSYSWKR	Q71QE8	*Crotalus viridis viridis*, phospholipase A_2
Y(MF)(YL)DFLCTDPSEKC	YMFYLDFLCTDPSEK	Q71QE8√	*C. viridis viridis*, phospholipase A_2
(IVS)WGGDI(CA)Q(PH)EPGVY(TK)	IVSWGGDICAQPHEPGHYTK	Q9I961	*Agkistrodon acutus*, Acubin2
	IVSWGGDPCAQPREPGVYTK	Q71QH8√	*T. stejnegeri*, serine protease CL4
	IVSWGGDICAQPREPEPYTK	Q2QA04	*C. durissus durissus*, serine proteinase

als for spectrum and contig coverage of all venom proteins).

Fig. 4 and Table I demonstrate that shotgun protein sequencing is a modification-tolerant approach applicable to protein mixtures. On the IKKβ protein, 100 different amino acids were found to be modified in at least one spectrum, and the whole dataset contained over a thousand spectra from hundreds of modified peptides. Nevertheless we were able to assemble 87% of all regions covered by at least three spectra and to derive *de novo* sequences that were found to be over 90% correct. Moreover we observed that errors predominantly fall into the initial/terminal regions of the contigs where there are fewer peaks to reliably call amino acids. Similar results were obtained on the venom dataset even though it contained almost 3000 different peptides from a mixture of *C. atrox* venom proteins. This 3.5-fold increase in the number of different peptides did not affect our sequencing accuracy and resulted in a 2-fold increase in the number of sequenced amino acids (IKKβ *versus* venom). Although the total length of all proteins identified on the venom dataset is ~4 times that of the IKKβ protein, much of the additional peptide diversity in the former is actually coming from the same protein regions. This is evidenced both by a larger number of peptides per contig and by the increase in sequencing coverage: more peptides per contig lead to an increased probability of finding spectrum peaks for all amino acids.

The majority of all contig sequences was readily identifiable as a peptide from the corresponding database (84% for the IKKβ dataset and 70% for the venom dataset). However, the latter also resulted in a significant number of contig sequences that did not match any proteins from the target species but had a significant match to other related species when matched against the database (using blastp (49) and SPIDER (50)). These are listed in Table II as *homologous* peptides and represent 14% of all *de novo* sequences obtained in the venom dataset (see supplemental materials for a complete list). As it turned out, for 19 of the 28 homologous contigs the assembled spectra could also be identified by database search (*i.e.* the peptide existed in a protein from a different species), and the found peptides matched our *de novo* sequence. On the remaining nine cases the assembled spectra did not match any peptide in the database, and thus this step neither confirmed nor refuted the putative homologies. All of these novel homologies were derived from contigs assembling multiple peptides where the annotated MS/MS spectra strongly supported the recovered sequences (see supplemental materials). It should also be noted that all *C. atrox* homologies were either matched to a different snake species or can be explained by single nucleotide polymorphisms of the original sequences, which were also detected in our sample. Together with the 13 homologous peptides that matched only venom proteins from other species, these results suggest that some *C. atrox* venom proteins still remain unknown. Moreover all homologous peptides were found among proteins from other snakes thus reinforcing our predictions.

In addition to homologous peptides, some contig sequences showed no similarity to any peptide in UniProtKB. Moreover these contigs contained only spectra that were not identified by traditional database search of the individual spectra. In the venom dataset, it turned out that six of 18 such unidentified contigs yielded highly reliable *de novo* sequences containing a long tag of 10 or more amino acids (allowing for one summed mass of two amino acids), thus again suggest-

Downloaded from http://www.mcponline.org/ by guest on September 15, 2019

TABLE III
Putative new C. atrox peptides with no homologous matches in Swiss-Prot/UniProtKB

Parentheses indicate portions where the order of the amino acids was not determined; square brackets indicate indistinguishable amino acid masses (on ion trap spectra); numbers in square brackets indicate mass intervals that could be explained by different amino acid compositions. The annotated MS/MS spectra for these contigs can be found in our supplemental materials.

De novo sequence	Number of assembled spectra
[Q,K]FGP[Q,K]NPFCF[I,L]VQK	7
QRAV[218.0][I,L]DEYPESVAHNF	5
(MT)TGDSE[I,L]SVCW	4
YWPNTD[Q,K]E[I,L]G[I,L]DK	5
AAYPWNPVASTTLCASAE[371.0]	10
[242.3]D[I,L]SED[Q,K]D[I,L][Q,K]AEVNK	3

ing a few still unknown proteins in *C. atrox* venom (see Table III for sequences and supplemental materials for annotated MS/MS spectra).

A small number of the assembled contigs turned out to be incorrect (due to incorrect alignments of spectra from different peptides) or to yield mostly incorrect *de novo* sequences that did not match the peptide sequences assigned to the assembled spectra by traditional database search. These were mostly caused by spuriously matching both *b* and *y* peaks or high intensity unexplained peaks in the assembled spectra and account for less than 5% of all assembled contigs.

DISCUSSION

Shotgun protein sequencing is a modification-tolerant approach to the interpretation of tandem mass spectra that enables *de novo* sequencing of protein mixtures even on ion trap instruments. Our approach, for the first time, demonstrates the feasibility of very accurate *de novo* sequencing of modified proteins into contigs (20 amino acids and longer) covering contiguous sequence regions up to 108 amino acids long. In fact, the extensive contig coverage of all regions with three or more overlapping peptides indicates that the major difficulty preventing the assembly of whole proteins is the strong bias in proteolytic digestion. Thus, one straightforward route toward the production of longer contigs is through the generation of richer peptide ladders using proteases with diminished cleavage specificity. Indeed the coverage observed in the venom dataset (based on a slightly improved digestion protocol) is already much better than the fragmented coverage of IKKβ (Fig. 4, compare *a* and *b*). In the context of deuterium exchange studies (30, 51), much progress has been achieved with controlled pepsin digests.

Using mass spectrometry for shotgun protein sequencing results in certain limitations that are without counterpart in the DNA sequencing realm. The sampling frequency of the amino acids across a protein sequence is not uniform and is dictated by local sequence context. The coverage of a protein by its peptides is biased by the specificity and distribution of cleavage sites of the proteases used. The ionizability and extent of fragmentation of individual peptides are biased by the presence/absence of basic, charge-bearing residues (Arg, Lys, and His) and Pro, whose constrained side chain is covalently bound to the peptide backbone. Certain combinations of amino acids have identical elemental compositions that are indistinguishable by mass and may leave ambiguity in the draft (or even finished) sequences depending on the completeness of fragmentation in the MS/MS spectra (Ile = Leu = 113, GG = Asn = 114, and GA = Gln = 128). Others have the same nominal mass but not elemental composition and are distinguishable only in MS/MS from high resolution instruments (Gln = Lys = 128 and Trp = DA = VS = 186). Distinguishing the identical elemental composition of isoleucine and leucine may be achievable by performing MS*n* to further fragment the Ile/Leu-specific immonium ion at *m/z* 86 (52) or, to a limited extent, by capitalizing on the cleavage specificity of chymotrypsin.

High resolution mass spectrometers, such as the Thermo LTQ-Orbitrap, may seamlessly elevate shotgun protein sequencing to a whole new level of productivity. In principle, higher mass accuracy should be directly translatable into much more sensitive detection of overlaps between spectra with poor *b/y* ion ladders. This increased sensitivity would be particularly relevant for the case of MS/MS spectra from highly charged (3+) peptides, which usually feature poor *b/y* ion fragmentation; these peptides tend to span more than one contig and could thus serve as "connectors" between adjacent contigs. Also when LC time scale-compatible electron transfer dissociation (53) becomes available, CNBr-derived long peptides may yield near complete, contiguous sequences.

Nonetheless even with a standard experimental setup and using only a relatively small MS/MS dataset from a modest resolution mass spectrometer, our approach very rapidly generated much more information about western diamondback rattlesnake venom proteins than some of the more laborious Edman degradation/cloning studies (36). Moreover these contigs can be easily produced with minimal experimental and computational effort whereas Edman degradation projects often take months to complete. Furthermore our contigs may be readily aligned and ordered by comparative protein sequencing that, akin to comparative DNA sequencing, utilizes previously determined protein sequences from evolutionarily close species. For example, one can use the *Crotalus durissus durissus* catrocollastatin protein sequence to map and order our *C. atrox* catrocollastatin contigs and obtain long sequences up to 96 amino acids in length.

Although defining the termini of mature proteins could be accomplished by using amine- and carboxyl-reactive labeling agents prior to enzymatic digestion, determining the signal peptides that are post-translationally cleaved would require gene cloning. To this end, the readily available contigs can be used to design degenerate DNA primers/probes to enable

Downloaded from http://www.mcponline.org/ by guest on September 15, 2019

subsequent gene cloning efforts from venom gland cDNA libraries.

Acknowledgments—We thank Dekel Tsur, Ari Frank, Stephen Tanner, Vineet Bafna, and Ebrahim Zandi for insightful discussions.

* This work was supported by NIGMS, National Institutes of Health Grant 1-R01-RR16522. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

S The on-line version of this article (available at http://www.mcponline.org) contains supplemental material.

§ To whom correspondence should be addressed: Dept. of Computer Science and Engineering, University of California San Diego, 9500 Gilman Dr., La Jolla, CA 92093. E-mail: bandeira@cs.ucsd.edu.

REFERENCES

1. Gearhart, P. J. (2002) Immunology: the roots of antibody diversity. *Nature* **419,** 29–31
2. Wiles, M. and Andreassen, P. (2006) Monoclonals—the billion dollar molecules of the future. *Drug Discov. World* **Fall,** 17–23
3. Haurum, J. S. (2006) Recombinant polyclonal antibodies: the next generation of antibody therapeutics? *Drug Discov. Today* **11,** 655–660
4. Lewis, R. J., and Garcia, M. L. (2003) Therapeutic potential of venom peptides. *Nat. Rev. Drug Discov.* **2,** 790–802
5. Pimenta, A. M., and De Lima, M. E. (2005) Small peptides, big world: biotechnological potential in neglected bioactive peptides from arthropod venoms. *J. Pept. Sci.* **11,** 670–676
6. Joseph, J. S., and Kini, R. M. (2004) Snake venom prothrombin activators similar to blood coagulation factor Xa. *Curr. Drug Targets Cardiovasc. Haematol. Disord.* **4,** 397–416
7. Swenson, S., Toombs, C. F., Pena, L., Johansson, J., and Markland, F. S. (2004) α-Fibrinogenases. *Curr. Drug Targets Cardiovasc. Haematol. Disord.* **4,** 417–435
8. Kini, R. M., Rao, V. S., and Joseph, J. S. (2001) Procoagulant proteins from snake venoms. *Haemostasis* **31,** 218–224
9. Swenson, S., Costa, F., Minea, R., Sherwin, R. P., Ernst, W., Fujii, G., Yang, D., and Markland, F. S. (2004) Intravenous liposomal delivery of the snake venom disintegrin contortrostatin limits breast cancer progression. *Mol. Cancer Ther.* **3,** 499–511
10. Pal, S. K., Gomes, A., Dasgupta, S. C., and Gomes, A. (2002) Snake venom as therapeutic agents: from toxin to drug development. *Indian J. Exp. Biol.* **40,** 1353–1358
11. Markland, F. S., Shieh, K., Zhou, Q., Golubkov, V., Sherwin, R. P., Richters, V., and Sposto, R. (2001) A novel snake venom disintegrin that inhibits human ovarian cancer dissemination and angiogenesis in an orthotopic nude mouse model. *Haemostasis* **31,** 183–191
12. Zugasti-Cruz, A., Maillo, M., López-Vera, E., Falcón, A., Heimer de la Cotera, E. P., Olivera, B. M., and Aguilar, M. B. (2006) Amino acid sequence and biological activity of a γ-conotoxin-like peptide from the worm-hunting snail Conus austini. *Peptides* **27,** 506–511
13. Ogawa, Y., Yanoshita, R., Kuch, U., Samejima, Y., and Mebs, D. (2004) Complete amino acid sequence and phylogenetic analysis of a long-chain neurotoxin from the venom of the African banded water cobra, Boulengerina annulata. *Toxicon* **43,** 855–858
14. Buczek, O., Bulaj, G., and Olivera, B. M. (2005) Conotoxins and the post-translational modification of secreted gene products. *Cell. Mol. Life Sci.* **62,** 3067–3079
15. Pimenta, A. M., Rates, B., Bloch, C., Gomes, P. C., Santoro, M. M., de Lima, M. E., Richardson, M., and Cordeiro, M. d. N. (2005) Electrospray ionization quadrupole time-of-flight and matrix-assisted laser desorption/ionization tandem time-of-flight mass spectrometric analyses to solve micro-heterogeneity in post-translationally modified peptides from Phoneutria nigriventer (Aranea, Ctenidae) venom. *Rapid Commun. Mass Spectrom.* **19,** 31–37
16. Daltry, J. C., Wüster, W., and Thorpe, R. S. (1996) Diet and snake venom evolution. *Nature* **379,** 537–540
17. Menezes, M. C., Furtado, M. F., Travaglia-Cardoso, S. R., Camargo, A. C., and Serrano, S. M. (2006) Sex-based individual variation of snake venom proteome among eighteen Bothrops jararaca siblings. *Toxicon* **47,** 304–312
18. Dos-Santos, M. C., Assis, E. B., Moreira, T. D., Pinheiro, J., and Fortes-Dias, C. L. (2005) Individual venom variability in Crotalus durissus ruruima snakes, a subspecies of Crotalus durissus from the Amazonian region. *Toxicon* **46,** 958–961
19. Escoubas, P. (2006) Mass spectrometry in toxinology: a 21st-century technology for the study of biopolymers from venoms. *Toxicon* **47,** 609–613
20. Fox, J. W., Ma, L., Nelson, K., Sherman, N. E., and Serrano, S. M. (2006) Comparison of indirect and direct approaches using ion-trap and Fourier transform ion cyclotron resonance mass spectrometry for exploring viperid venom proteomes. *Toxicon* **47,** 700–714
21. Soares, M. R., Oliveira-Carvalho, A. L., Wermelinger, L. S., Zingali, R. B., Ho, P. L., Junqueira-de Azevedo, I. d. L., and Diniz, M. R. (2005) Identification of novel bradykinin-potentiating peptides and C-type natriuretic peptide from Lachesis muta venom. *Toxicon* **46,** 31–38
22. Wermelinger, L. S., Dutra, D. L., Oliveira-Carvalho, A. L., Soares, M. R., Bloch, C., and Zingali, R. B. (2005) Fast analysis of low molecular mass compounds present in snake venom: identification of ten new pyroglutamate-containing peptides. *Rapid Commun. Mass Spectrom.* **19,** 1703–1708
23. Fischer, B., Roth, V., Roos, F., Grossmann, J., Baginsky, S., Widmayer, P., Gruissem, W., and Buhmann, J. M. (2005) Novohmm: a hidden Markov model for de novo peptide sequencing. *Anal. Chem.* **77,** 7265–7273
24. Chen, T., Kao, M. Y., Tepel, M., Rush, J., and Church, G. M. (2001) A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry. *J. Comput. Biol.* **8,** 325–337
25. Dancík, V., Addona, T. A., Clauser, K. R., Vath, J. E., and Pevzner, P. A. (1999) De novo peptide sequencing via tandem mass spectrometry. *J. Comput. Biol.* **6,** 327–342
26. Johnson, R. S., and Biemann, K. (1987) The primary structure of thioredoxin from Chromatium vinosum determined by high-performance tandem mass spectrometry. *Biochemistry* **26,** 1209–1214
27. Pham, V., Henzel, W. J., Arnott, D., Hymowitz, S., Sandoval, W. N., Truong, B. T., Lowman, H., and Lill, J. R. (2006) De novo proteomic sequencing of a monoclonal antibody raised against ox40 ligand. *Anal. Biochem.* **352,** 77–86
28. Bandeira, N., Tang, H., Bafna, V., and Pevzner, P. (2004) Shotgun protein sequencing by tandem mass spectra assembly. *Anal. Chem.* **76,** 7221–7233
29. Klammer, A. A., and MacCoss, M. J. (2006) Effects of modified digestion schemes on the identification of proteins from complex mixtures. *J. Proteome Res.* **5,** 695–700
30. Englander, J. J., Del Mar, C., Li, W., Englander, S. W., Kim, J. S., Stranz, D. D., Hamuro, Y., and Woods, V. L. (2003) Protein structure change studied by hydrogen-deuterium exchange, functional labeling, and mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 7057–7062
31. MacCoss, M. J., McDonald, W. H., Saraf, A., Sadygov, R., Clark, J. M., Tasto, J. J., Gould, K. L., Wolters, D., Washburn, M., Weiss, A., Clark, J. I., and Yates, J. R. (2002) Shotgun identification of protein modifications from protein complexes and lens tissue. *Proc. Natl. Acad. Sci. U. S. A.,* **99,** 7900–7905
32. Vinson, 32 J. P., Jaffe, D. B., O'Neill, K., Karlsson, E. K., Stange-Thomann, N., Anderson, S., Mesirov, J. P., Satoh, N., Satou, Y., Nusbaum, C., Birren, B., Galagan, J. E., and Lander, E. S. (2005) Assembly of polymorphic genomes: algorithms and application to Ciona savignyi. *Genome Res.* **15,** 1127–1135
33. Bandeira, N., Tsur, D., Frank, A., and Pevzner, P. A. (2006) A new approach to protein identification, in *Proceeding of the Tenth Annual International Conference in Research in Computational Molecular Biology (RECOMB 2006)* (Apostolico, A., Guerra, C., Istrail, S., Pevzner, P. A., and Waterman, M., eds) Vol. 3909, pp. 363–378, Springer, Germany
34. Bandeira, N., Tsur, D., Frank, A., and Pevzner, P. A. (2007) Protein identification by spectral networks analysis. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 6140–6145
35. Pevzner, P. A., Tang, H., and Tesler, G. (2004) De novo repeat classification and fragment assembly. *Genome Res.* **14,** 1786–1796
36. Zhou, Q., Smith, J. B., and Grossman, M. H. (1995) Molecular cloning and expression of catrocollastatin, a snake venom protein from Crotalus atrox (western diamondback rattlesnake) which inhibits platelet adhesion

Downloaded from http://www.mcponline.org/ by guest on September 15, 2019

to collagen. *Biochem. J.* **307,** 411–417

37. Miller, B. S., and Zandi, E. (2001) Complete reconstitution of human IκB kinase (IKK) complex in yeast. Assessment of its stoichiometry and the role of IKKγ on the complex activity in the absence of stimulation. *J. Biol. Chem.* **276,** 36320–36326
38. Tanner, S., Shu, H., Frank, A., Wang, L. C., Zandi, E., Mumby, M., Pevzner, P. A., and Bafna, V. (2005) InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Anal. Chem.* **77,** 4626–4639
39. Hu, M. C., and Hung, M. C. (2005) Role of IκB kinase in tumorigenesis. *Future Oncol.* **1,** 67–78
40. Tsur, D., Tanner, S., Zandi, E., Bafna, V., and Pevzner, P. A. (2005) Identification of post-translational modifications by blind search of mass spectra. *Nat. Biotechnol.* **23,** 1562–1567
41. Geoghegan, K. F., Hoth, L. R., Tan, D. H., Borzilleri, K. A., Withka, J. M., and Boyd, J. G. (2002) Cyclization of N-terminal S-carbamoylmethylcysteine causing loss of 17 Da from peptides and extra peaks in peptide maps. *J. Proteome Res.* **1,** 181–187
42. Frank, A., and Pevzner, P. (2005) PepNovo: de Novo peptide sequencing via probabilistic network modeling. *Anal. Chem.* **77,** 964–973
43. Smith, T. F., and Waterman, M. S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.* **147,** 195–197
44. Leitner, A., and Lindner, W. (2004) Current chemical tagging strategies for proteome analysis by mass spectrometry. *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.* **813,** 1–26
45. Guerrera, I. C., and Kleiner, O. (2005) Application of mass spectrometry in proteomics. *Biosci. Rep.* **25,** 71–93
46. Savitski, M. M., Nielsen, M. L., Kjeldsen, F., and Zubarev, R. A. (2005) Proteomics-grade de novo sequencing approach. *J. Proteome Res.* **4,** 2348–2354
47. Myers, E. W. (1995) Toward simplifying and accurately formulating fragment assembly. *J. Comput. Biol.* **2,** 275–290
48. Bartels, C. (1990) Fast algorithm for peptide sequencing by mass spectroscopy. *Biomed. Environ. Mass Spectrom.* **19,** 363–368
49. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402
50. Han, Y., Ma, B., and Zhang, K. (2005) SPIDER: software for protein identification from sequence tags with de novo sequencing error. *J. Bioinform. Comput. Biol.* **3,** 697–716
51. Pantazatos, D., Kim, J. S., Klock, H. E., Stevens, R. C., Wilson, I. A., Lesley, S. A., and Woods, V. L. (2004) Rapid refinement of crystallographic protein construct definition employing enhanced hydrogen/deuterium exchange MS. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 751–756
52. Armirotti, A., Millo, E., and Damonte, G. (2007) How to discriminate between leucine and isoleucine by low energy ESI-trap MSn. *J. Am. Soc. Mass Spectrom.* **18,** 57–63
53. Syka, J. E., Coon, J. J., Schroeder, M. J., Shabanowitz, J., and Hunt, D. F. (2004) Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 9528–9533
54. Pevzner, P. A., Dancík, V., and Tang, C. L. (2000) Mutation-tolerant protein identification by mass spectrometry. *J. Comput. Biol.* **7,** 777–787

Downloaded from http://www.mcponline.org/ by guest on September 15, 2019