

Shotgun Protein Sequencing of Post-translationally Modified Snake Venom Proteins

Nuno Bandeira*

Karl Clauser[†]

Pavel A. Pevzner*

Abstract

Although snake venom proteins have been proven instrumental in the design of blood clotting and cytostatic breast cancer drugs, the main method in use to sequence these unknown proteins is still the low-throughput process of Edman degradation. Moreover, the common approach to MS/MS protein sequence identification is limited in that it focuses on the identification of single MS/MS spectra from mostly non-overlapping peptides, insufficient for whole protein reconstruction. Having shown that, for samples containing only purified proteins, large portions of the protein sequence can be recovered by combining multiple MS/MS spectra from overlapping peptides, we now extend the Shotgun Protein Sequencing approach to the analysis of mixtures of modified proteins. A mixture of proteins from *crotalus atrox* (western diamondback rattlesnake) venom was analyzed using LC/MS/MS and the correct overlaps between MS/MS spectra reliably identified. These overlaps were then assembled into multiple alignments and large portions of their amino acid sequences accurately reconstructed by de-novo interpretation.

On more than one occasion, proteins from snake venom have provided essential clues for the design of important drugs [1] - examples include drugs for controlling blood coagulation and drugs for breast and ovarian cancer treatment. Despite the vital importance of venom proteins, the DNA sequences of the various snake genomes are neither known nor likely to become available anytime soon. Thus, the mainstream method to sequence these unknown proteins is still the restrictive and low-throughput procedure of Edman degradation [2] - a task made difficult by protein purification procedures, unexpected post-translational modifications and naturally blocked protein N-termini. These problems gain additional relevance when one considers limited amounts of venom protein extracts and the known high level of post-translational modifications expected to occur on venom proteins. These difficulties have been widely acknowledged and have motivated several attempts at de-novo sequencing of tandem mass spectrometric (MS/MS) data from venom proteins [2]. However, all such attempts were made using traditional de-novo sequencing approaches that consider each MS/MS spectrum in isolation and thus face multiple difficulties in the reliable assignment of peptide sequences to MS/MS spectra.

Conceptually, sequencing a protein from a set of MS/MS spectra is a simple procedure that can be described by an easy analogy. Imagine you have a container 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 inter-bead distances is initially chosen by the designer and all necklaces are then made using exactly the same specifications. Now assume that one day you open your necklace 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 bead necklaces, as specified by the complete set of consecutive inter-bead distances? The correspondence between this allegory and protein sequences is straightforward: inter-bead distances correspond to amino acid masses and beads correspond to backbone fragmentation points (between consecutive amino acids).

Even though the intrinsic characteristics of MS/MS data add more than a few difficulties to this assembly process, as far back as 1989, Hopper et al. [3] recognized the potential of tandem mass spectrometry for protein sequencing and manually sequenced a complete protein from the rabbit bone marrow. With the same purpose in mind we have previously introduced the Shotgun Protein Sequencing approach [4] that significantly improves the quality and extent of de-novo protein sequence reconstruction by combining multiple MS/MS spectra from overlapping peptides (these can nowadays be easily generated using non-specific proteases or several different proteases with different specificities [5]). While this approach proved to be efficient for the assembly of purified unmodified proteins, it was not designed to meet the challenges posed by mixtures of modified proteins. We have thus redesigned our Shotgun Protein Sequencing approach to overcome these challenges and enable simultaneous sequencing of several unknown proteins in the same sample, while allowing for unknown or unexpected modifications on some or all proteins. As illustrated in Figure 1, our approach consists of three sequential stages. First we separate *b/y*-ions using *spectral pairs* (pairs of spectra obtained from overlapping peptides or from unmodified and modified versions of the same peptide [6]) by

*University of California, San Diego, Dept. of Computer Science and Engineering, 9500 Gilman Drive, La Jolla, CA 92093, USA.[†]Broad Institute of MIT and Harvard, Cambridge, MA.

selecting peaks on the adequate diagonal. Thus, the expected input to our second stage is a set of pairs of spectra where each spectrum contains only *b*-ions. We then interpret each spectrum as a set of vertices (one per spectrum peak) and merge vertices from different spectra if the corresponding peaks were matched in the spectral pair. Furthermore, vertices are connected by an edge if the corresponding spectrum peaks differ by an amino acid mass. In the final interpretation stage we simply read the amino acid sequence from the resulting graph. Amino acid sequences recovered by this procedure have the sequencing statistics shown in Figure 2.

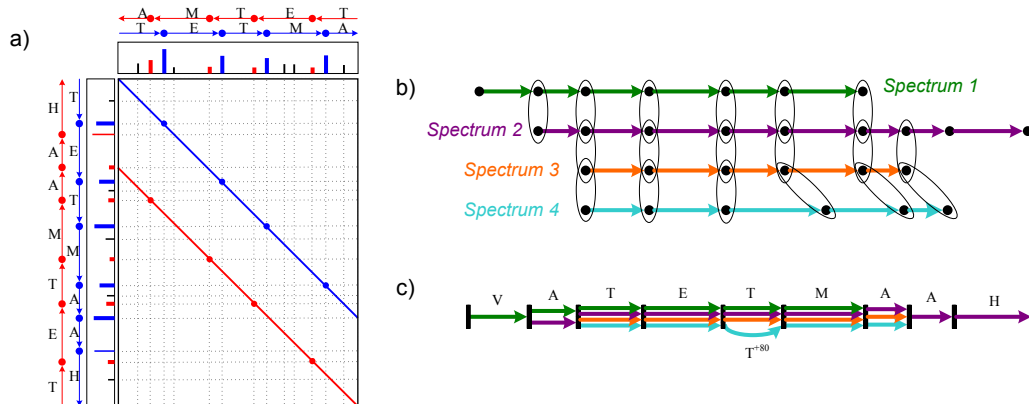


Figure 1: Shotgun Protein Sequencing assembly procedure: **a)** Identify spectral pairs and use them to separate *b*/*y*-ions; *spectral pairs* are pairs of spectra obtained from overlapping peptides or from unmodified and modified versions of the same peptide [6]. **b)** Collapse all matched ions into single vertices; two vertices are connected by an edge if the corresponding ions differ by an amino acid mass. **c)** Read the amino acid sequence from the resulting graph.

IKKb sample	Whole-protein sequencing accuracy
IKKb	94%
Endoproteinase Glu-C	97%
Trypsin	94%
Overall (12 proteins)	92%
Crotalus atrox venom sample: top 5 most abundant proteins	
Catrocollastatin precursor	90%
Hemorrhagic metalloproteinase HT-E precursor	87%
Vascular apoptosis-inducing protein 1	99%
Phospholipase A2 homolog Cax-K49	92%
Phospholipase A2 precursor	94%
Overall (33 proteins)	90%

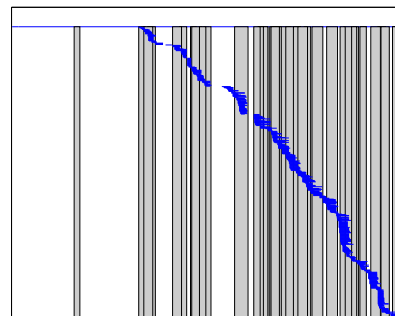


Figure 2: Experimental results. The table on the left shows the observed amino acid prediction accuracy on two different datasets. The figure on the right illustrates the sequencing coverage of the most abundant venom protein (Catrocollastatin precursor, 609 amino acids long); the line at the top represents the target protein sequence, each blue horizontal line represents one identified peptide, areas sequenced by our approach are shown in gray.

References

- [1] A M Pimenta and M E De Lima. Small peptides, big world: biotechnological potential in neglected bioactive peptides from arthropod venoms. *J Pept Sci*, 11:670–676, 2005.
- [2] P Escoubas. Mass spectrometry in toxinology: A 21st-century technology for the study of biopolymers from venoms. *Toxicon*, 47:609–613, 2006.
- [3] S. Hopper, RS. Johnson, JE. Vath, and K. Biemann. Glutaredoxin from rabbit bone marrow. Purification, characterization, and amino acid sequence determined by tandem mass spectrometry. *J Biol Chem*, 264(34):20438–20447, 1989.
- [4] N. Bandeira, H. Tang, V. Bafna, and P. Pevzner. Shotgun protein sequencing by tandem mass spectra assembly. *Analytical Chemistry*, 76:7221–7233, 2004.
- [5] A A Klammer and M J MacCoss. Effects of modified digestion schemes on the identification of proteins from complex mixtures. *J Proteome Res*, 5:695–700, 2006.
- [6] N. Bandeira, D. Tsur, A. Frank, and P.A. Pevzner. A New Approach to Protein Identification. In Apostolico A., Guerra C., Istrail S., Pevzner P.A., and Waterman M., editors, *Proceeding of the Tenth Annual International Conference in Research in Computational Molecular Biology (RECOMB 2006)*, volume 3909 of *Lecture Notes in Computer Science*, pages 363 – 378, 2006.