Modification-tolerant Shotgun Protein Sequencing of a Snake Venom Proteome

Nuno Bandeira[1], Karl Clauser[2], Pavel Pevzner[1]

Keywords: Protein sequencing, De novo, Assembly, Tandem Mass Spectrometry

1 Introduction.

Despite the steady accumulation of fully sequenced genomes for model organisms, limited or no sequence information is available for most organisms. Moreover, natural mechanisms of variation such as accelerated mutation and combinatorial recombination in immunoglobulins regularly create novel sequences in the proteomes of model organisms. However, since protein identification via database searching of MS/MS spectra is not applicable in these contexts and de novo sequencing of individual peptides by MS/MS tends to produce inaccurate short sequences, laborious Edman degradation of isolated peptides remains the primary approach to design degenerate PCR primers for cloning the genes of individual proteins. To catalyze the otherwise impractical application of proteomics methodologies in these cases, we have developed Shotgun Protein Sequencing to produce highly accurate *protein contigs* by acquiring, assembling and simultaneously sequencing multiple MS/MS spectra from overlapping (and possibly modified) peptides generated via multiple proteases of different specificities. By capitalizing on the redundant sequence information in multiple spectra covering the same protein regions, we were able to achieve substantially higher de novo sequencing accuracy than what is achievable through the interpretation of individual spectra. Thus we believe that Shotgun Protein Sequencing opens the door to automated proteomic studies in otherwise neglected organisms that lack sequenced genomes.

2 Method and Results.

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, Klaus Biemann's group [2] 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 developed

[1]Department of Computer Science and Engineering, University of California, San Diego, CA, USA. E-mails: bandeira@cs.ucsd.edu, ppevzner@cs.ucsd.edu

[2]Broad Institute of MIT and Harvard, Cambridge, MA, USA. E-mail: clauser@broad.mit.edu

the Shotgun Protein Sequencing approach that enables simultaneous sequencing of several unknown proteins in the same sample while allowing for unexpected modifications on some or all proteins (see Figure 1). In this approach, a set of assembled spectra is called a *protein contig* and all spectra in a contig are simultaneously interpreted to generate a single de novo sequence. Our algorithm was benchmarked on a sample from inhibitor of nuclear factor kappa B kinase beta (IKKb) and on venom from western diamondback rattlesnakes, which is a mixture of about two dozen proteins. On the latter, we obtained contig sequences covering 87% of all protein regions covered by 3 or more spectra and generated de novo sequences at an average amino acid prediction accuracy of 90%, even on low-accuracy Ion Trap MS/MS spectra. In addition to resequencing large portions of the known western diamondback rattlesnake venom proteins, Shotgun Protein Sequencing identified a) several polymorphisms where both the reference and homologous variant were present in our dataset and b) putative novel sequences with strong homology to known venom proteins from other snake species. The observation of these novel peptides with strong homology to venom proteins from related snake species readily suggests that venom proteins are more diverse than is currently known and both advocates the need for and applicability of novel MS-based protein sequencing approaches.

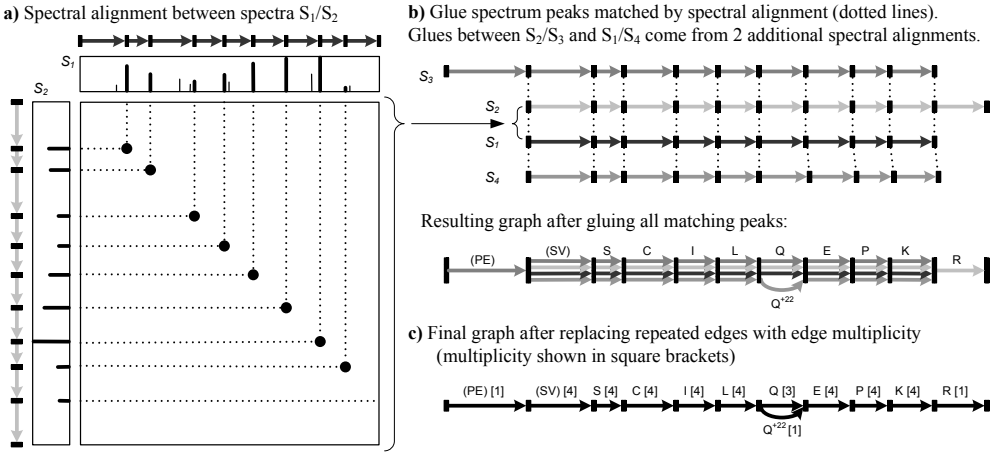

Figure 1: Shotgun Protein Sequencing through assembly of tandem mass spectra; **a)** Spectral alignment between spectrum S_1 (from peptide SVSCILQEPK) and spectrum S_2 (from peptide SVSCILQEPKR) reveals the common sequence information in both spectra. **b)** Matching peaks in spectral alignments become pairwise gluing instructions between every pair of aligned spectra. Additional spectra S_3 (from PESVSCILQEPK) and S_4 (from SVSCILQ^{+22}EPK) respectively illustrate additional types of spectral alignment: partial peptide overlap and alignment of modified/unmodified variants of the same peptide [1]; **c)** Repeated edges are replaced by single edges with weight proportional to their multiplicity and the consensus sequence for all assembled spectra is found by the heaviest path in this graph.

References

[1] N. Bandeira, D. Tsur, A. Frank, and P.A. Pevzner. 2006 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)*, pp. 363–378.

[2] S. Hopper, RS. Johnson, JE. Vath, and K. Biemann. 1989 Glutaredoxin from rabbit bone marrow. Purification, characterization, and amino acid sequence determined by tandem mass spectrometry. *J Biol Chem*, 264:20438–20447.