Overlapping MS/MS spectra and disease proteomics

Nuno Bandeira

Department of Computer Science and Engineering,
University of California, San Diego,
9500 Gilman Drive, La Jolla, CA 92093, USA

The ongoing success of the proteomics endeavor is the result of a prolific symbiosis between experimental ingenuity [2, 3, 4] and efficient bioinformatics [5, 6, 7, 8, 9, 10, 11]. Without these, ground-breaking landmarks such as the human genome project [12, 13] or the HUPO initiative [14] would likely not have seen the light of day. But despite valuable contributions, the road to a better understanding of disease proteomics is still hurdled by significant difficulties in the extensive identification of post-translational modifications and in the sequencing of novel proteins like cancer fusion proteins or antibody chains.

Recently, tandem mass spectrometry (MS/MS) based approaches seemed to be reaching the limit on the amount of information that could be extracted from MS/MS spectra [15, 16, 17]. However, a closer look reveals that a common limiting procedure is to analyze each spectrum in isolation, even though high throughput mass spectrometry regularly generates many spectra from related peptides.

By capitalizing on this redundancy we have shown that, similarly to the alignment of protein sequences [5], unidentified MS/MS spectra can also be aligned for the identification of modified and unmodified variants of the same peptide. Moreover, this alignment procedure can be iterated for the accurate grouping of multiple peptide variants (Figure 1). The highly correlated peaks in spectra from variants of the same peptide allowed us to reliably identify all known and even some unknown modifications in a sample of cataractous lenses proteins [18, 19].

Furthermore, the combination of shotgun proteomics [20] with the alignment of spectra from overlapping peptides led us to the development of Shotgun Protein Sequencing [21] - similarly to the assembly of DNA reads into whole genomic sequences, we have shown that assembly of MS/MS spectra enables the highest ever de-novo sequencing accuracy, while recovering over 85% of the target proteins sequence[1] [22](Figure 2). Similar mixtures of venom proteins have previously provided essential clues for the design of important drugs [23, 24].

Beyond providing the proof-of-concept for these methods, we are actively collaborating on quantifying drug and age-induced changes in post-translational modifications, and on sequencing of cancer fusion proteins, antibody light/heavy chains and unknown snake venom proteins. Additionally, our tools will be available to the community as open-source packages and web services [2].

[1] Covered by at least 3 overlapping spectra.

[2] Soon to be introduced to a wide audience at an invited tutorial affiliated with the Computational Systems Bioinformatics (CSB'2006) conference in Stanford, USA.

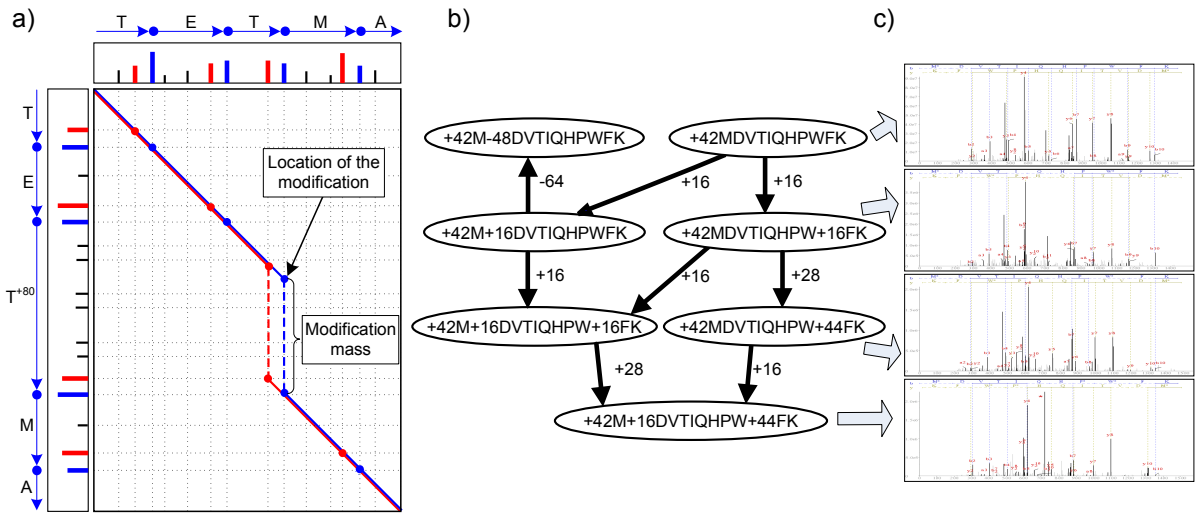
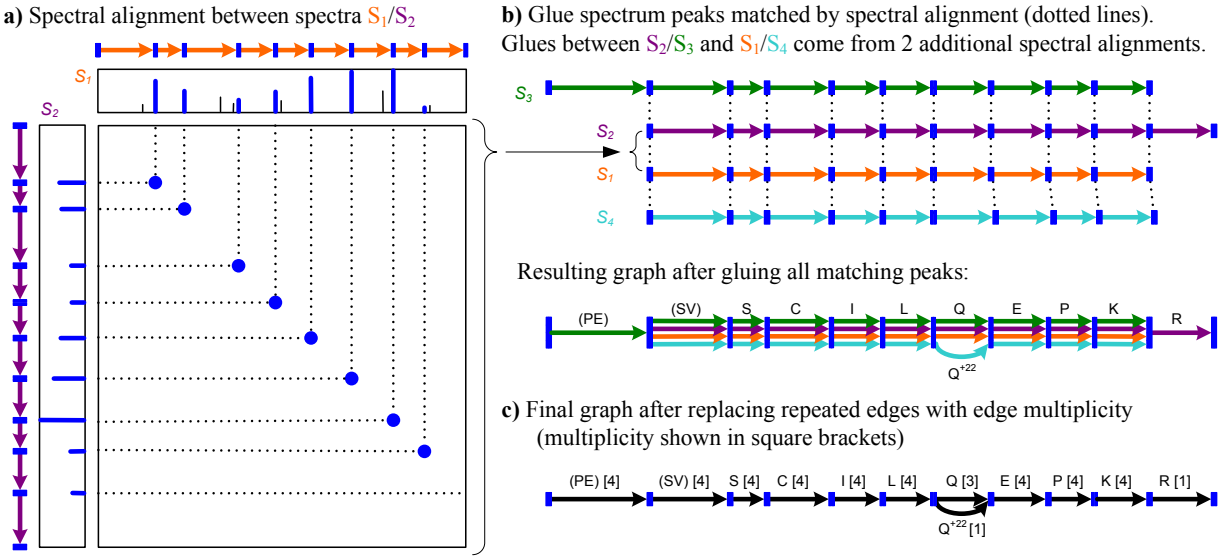


Figure 1: Identification of post-translational modifications through spectral alignment; **a)** Spectral alignment between modified and unmodified variants of the peptide TETMA (b-ions shown in blue, y-ions in red); **b)** Grouped modification states of the peptide MDVTIQHPWFK from a sample of cataractous lenses; **c)** Highly correlated MS/MS spectra from the indicated peptide variants.



d) Sequenced portions of the target protein sequence:

MSWSPSLTTQTCGAWEMKERLGTGGFGNVIRWHNQETGEQIAIKQCRQELSPRNRERWCLEIQIMRRLTHPNVVAARDVPEGMQNLAPNDLPLLAM
EYCQGGDLRKYLNQFENCCGLREGAILTLLSDIASALRYLHENRIIHRDLKPENIVLQQGEQRLIHKIIDLGYAKELDQGSLCTSFVGTLQYLAPE
LLEQQKYTVTVDYWSFGTLAFECITGFRPFLPNWQPVQWHSKVRQKSEVDIVVSEDLNGTVKFSSSLPYPNNLNSVLAERLEKWLQLMLMWHPRQR
GTDPTYGPNGCFKALDDILNLKLVHILNMVTGTIHTYPVTEDESLQSLKARIQQDTGIPEEDQELLQEAGLALIPDKPATQCISDGKLNEGHTLDM
DLVFLFDNSKITYETQISPRPQPESVSCILQEPKRNLAFFQLRKVWGQVWHSIQTLKEDCNRLQQGQRAAMMNLLRNNSCLSKMKNSMASMSQQLK
AKLDFFKTSIQIDLEKYSEQTEFGITSDKLLLAWREMEQAVELCGRENEVKLLVERMMALQTDIVDLQRSPMGRKQGGTLDDLEEQARELYRRLRE
KPRDQRTEGDSQEMVRLLLQAIQSFEKKVRVIYTQLSKTVVCKQKALELLPKVEEVVSLMNEDEKTVVRLQEKRQKELWNLLKIACSKVRGPVSGS
PDSMNASRLSQPGQLMSQPSTASNSLPEPAKKSEELVAEAHNLCTLLENAIQDTVREQDQSFTALDWSWLQTEEEEHSCLEQAS

Figure 2: 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; **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; **d)** Recovered portions of a target protein in our sample [22]. Correct amino acid predictions are shown in red (93%) and incorrect in blue (7%).

References

[1] N O Ku and M B Omary. A disease- and phosphorylation-related nonmechanical function for keratin 8. *J Cell Biol*, 174:115–125, 2006.

[2] A. Görg. Two-dimensional electrophoresis. *Nature*, pages 545–546, 1991.

[3] SP. Gygi, B. Rist, SA. Gerber, F. Turecek, MH. Gelb, and R. Aebersold. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol*, 17(10):994–999, 1999.

[4] M P Washburn, D Wolters, and J R Yates. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol*, 19:242–247, 2001.

[5] T F Smith and M S Waterman. Identification of common molecular subsequences. *J Mol Biol*, (1):195–197, 1981.

[6] JR. Yates, JK. Eng, and AL. McCormack. Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. *Anal Chem*, 67(18):3202–3210, 1995.

[7] S F Altschul, T L Madden, A A Schäffer, J Zhang, Z Zhang, W Miller, and D J Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, 25:3389–3402, 1997.

[8] V. Dancík, TA. Addona, KR. Clauser, JE. Vath, and PA. Pevzner. De novo peptide sequencing via tandem mass spectrometry. *J Comput Biol*, 6(3-4):327–342, Fall-Winter 1999.

[9] C H Wu, L S Yeh, H Huang, L Arminski, J Castro-Alvear, Y Chen, Z Hu, P Kourtesis, R S Ledley, B E Suzek, C R Vinayaka, J Zhang, and W C Barker. The protein information resource. *Nucleic Acids Res*, 31:345–347, 2003.

[10] AI. Nesvizhskii, A. Keller, E. Kolker, and R. Aebersold. A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem*, 75(17):4646–4658, 2003.

[11] R Apweiler, A Bairoch, C H Wu, W C Barker, B Boeckmann, S Ferro, E Gasteiger, H Huang, R Lopez, M Magrane, M J Martin, D A Natale, C O'Donovan, N Redaschi, and L S Yeh. Uniprot: the universal protein knowledgebase. *Nucleic Acids Res*, 32:115–119, 2004.

[12] ES. Lander and et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, Feb 2001.

[13] JC. Venter and et al. The sequence of the human genome. *Science*, 291(5507):1304–1351, Feb 2001.

[14] S. Hanash, J. Celis, and J. Rossier. HUPO (Human Proteome Organization) 1st World Congress. 21-24 November 2002, Versailles, France. Abstracts. *Mol Cell Proteomics*, 1:651–752, 2002.

[15] R Craig and R C Beavis. Tandem: matching proteins with tandem mass spectra. *Bioinformatics*, 20:1466–1467, 2004.

[16] S. Tanner, H. Shu, A. Frank, LC. Wang, E. Zandi, M. Mumby, PA. Pevzner, and V. Bafna. InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Anal Chem*, 77(14):4626–4639, Jul 2005.

[17] B Fischer, V Roth, F Roos, J Grossmann, S Baginsky, P Widmayer, W Gruissem, and J M Buhmann. Novohmm: a hidden markov model for de novo peptide sequencing. *Anal Chem*, 77:7265–7273, 2005.

[18] 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.

[19] N. Bandeira, D. Tsur, A. Frank, and P.A. Pevzner. Protein Identification via Spectral Networks Analysis. *(submitted)*, 2006.

[20] MJ. MacCoss, WH. McDonald, A. Saraf, R. Sadygov, JM. Clark, JJ. Tasto, KL. Gould, D. Wolters, M. Washburn, A. Weiss, JI. Clark, and JR. Yates. Shotgun identification of protein modifications from protein complexes and lens tissue. *Proc Natl Acad Sci U S A*, 99(12):7900–7905, 2002.

[21] N. Bandeira, H. Tang, V. Bafna, and P. Pevzner. Shotgun protein sequencing by tandem mass spectra assembly. *Analytical Chemistry*, 76:7221–7233, 2004.

[22] N. Bandeira, K. Clauser, and P.A. Pevzner. Shotgun Protein Sequencing of Snake Venom Proteins. *(in preparation)*, 2006.

[23] R J Lewis and M L Garcia. Therapeutic potential of venom peptides. *Nat Rev Drug Discov*, 2:790–802, 2003.

[24] 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.