Thesis proposal outline

Aaron A. Klammer
Department of Genome Sciences
University of Washington

April 25, 2005

1. Research Plan

   Hypotheses:

   (a) An improved predictive model of mass spectrum peak intensity will provide insight into the complex chemistry of peptide fragmentation.

   (b) Such a model will gain predictive power by including the contribution of chemical and instrument noise to spectrum variability.

   (c) A realistic intensity model will also be useful for improving identification of unknown peptide fragmentation spectra, especially in conjunction with a sequence database search step.

   Specific Aims:

   (a) The design and implementation of a probabilistic graphical model of spectrum intensity incorporating current understanding of peptide fragmentation chemistry that is trainable from mass spectrometry data.

   (b) The experimental generation of a set of high-confidence spectrum-peptide associations from peptides with controlled composition.

   (c) The incorporation of the above intensity model into a sequence database search algorithm.

2. Background

   (a) Intensity modelling.
      i. Mobile proton model. [11]
      ii. Basic residue content. [8]
      iii. Decision tree intensity model. [2]
      iv. HMM spectrum-peptide associations. [**?**]

   (b) Noise modelling; spectrum peak height variability.
      i. Spectrum averaging. [10]
      ii. Spectrum similarity. [9]

   (c) Sequence database search algorithm.
      i. SEQUEST [3]
      ii. RADARS [4]
      iii. Mascot [7]
      iv. SVM approach. [1]
      v. PeptideProphet. [5]
      vi. ProteinProphet. [6]

3. Preliminary Findings

   (a) Charge state paper.

(b) Digestion paper.

(c) Adele's paper.

4. Methods of Procedure

(a) The design and implementation of a model of incorporating current understanding of peptide fragmentation. This model will be trained on mass spectrometry peak intensity data. Implementations might include:

i. A deterministic model that generates a single predicted spectrum for a given peptide, analogous to SEQUEST or the mobile proton model. To be worthwhile, this model must include additional free parameters learned from actual data, such as:

A. Information about probabilities of fragmentations occurring adjacent to specific amino acids.

B. Information about continous rate equations of the known chemical reactions that participate in peptide fragmentation.

ii. A noise model that generates a spectrum distribution from a single predicted spectrum. Such a model would account for instrument noise, ignoring the noise resulting from the specific peptide that generated the spectrum.

iii. A hidden Markov model in which the states are fragment ions, the transitions are fragmentation events, and the emissions are peak intensities. This model would generate a probability distribution for all possible fragmentation spectra given a peptide.

iv. A model that generates a spectrum probability distribution from a peptide, using some combination of 4(a)iii and 4(a)ii.

(b) The experimental generation of a set of high-confidence spectrum-peptide associations from peptides with controlled composition. These spectra, along with spectra from other sources, will be used to test the intensity model described above, along with data from other sources. This aim might involve, for example, the following laboratory experiments:

i. Synthesizing a series of peptides that differ by a single amino acid or amino acid pair or some other simple difference and determining the global effect on all peak intensities and noise levels.

ii. Determining noise levels for specific peptides or spectrum peaks over thousands of spectra, or when eluted at different concentrations or under different solvent conditions.

iii. Synthesizing specific peptides which incorporate features for which could help choose between competing models or for which a particular model break downs.

(c) The incorporation of one or several of the intensity models described above into a sequence database search algorithm. This might involve the following:

i. Incorporating the deterministicly generated theoretical single spectrum for a each peptide into a search algorithm, to be used in manner analogous to SEQUEST.
ii. Using the spectrum distribution generated by the HMM to determine the peptide with the highest posterior probability given the spectrum.
iii. Using the HMM probability function in the context of a Fisher-kernel similarity score.

(d) Fisher kernel.

5. Alternative Approaches

(a) Discrete chemical equilibrium model. (i.e. Mobile proton model)

(b) Heuristic model. (i.e. SEQUEST theoretical spectrum)

(c) SVM regression.

(d) Topographic SVM.

(e) Protein identification technology in general:

i. Stan's aptamer project.
ii. Western blot.

6. Significance

(a) The most immediate benefit of this research will be improved understanding of peptide fragmentation.

(b) The most obvious direct benefit of this research will be improved peptide identification. Peptide identification is useful for solving a broad array of biological problems.

Bibliography

[1] D. C. Anderson, W. Li, D. G. Payan, and W. S. Noble. A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide MS/MS spectra and SEQUEST scores. *Journal of Proteome Research*, 2(2):137–146, 2003.

[2] J. E. Elias, F. D. Gibbons, O. D. King, F. P. Roth, and S. P. Gygi. Intensity-based protein identification by machine learning from a library of tandem mass spectra. *Nature Biotechnology*, 22:214–219, 2004.

[3] J. K. Eng, A. L. McCormack, and J. R. Yates, III. 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, 1994.

[4] H. I. Field, D. Fenyo, and R. C. Beavis. Radars, a bioinformatics solution that automates proteome mass spectral analysis, optimises protein identification, and archives data in a relational database. *Proteomics*, 2:36–47, 2002.

[5] A. Keller, A. I. Nezvizhskii, E. Kolker, and R. Aebersold. Empirical statistical model to estimate the accuracy of peptide identification made by MS/MS and database search. *Analytical Chemistry*, 74:5383–5392, 2002.

[6] A. I. Nesvizhskii, A. Keller, E. Kolker, and R. Aebersold. A statistical model for identifying proteins by tandem mass spectrometry. *Analytical Chemistry*, 75:4646–4658, 2003.

[7] D. N. Perkins, D. J. C. Pappin, D. M. Creasy, and J. S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20:3551–3567, 1999.

[8] D. L. Tabb, Y. Huang, V. H. Wysocki, and J. R. Yates, III. Influence of basic residue content on fragment ion peak intensities in low-energy collision-induced dissociation spectra of peptides. *Analytical Chemistry*, 76:1243–48, 2004.

[9] D. L. Tabb, M. J. MacCoss, C. C. Wu, S. D. Anderson, and J. R. Yates, III. Similarity among tandem mass spectra from proteomic experiments: detection, significance, and utility. *Analytical Chemistry*, 75:2470–2477, 2003.

[10] J. D. Venable and J. R. Yates, III. Impact of ion trap tandem mass spectra variability on the identification of peptides. *Analytical Chemistry*, 76(10):2928–2937, 2004.

[11] Z. Zhang. Prediction of low-energy collision-induced dissociation spectra of peptides. *Analytical Chemistry*, 76:3908–3922, 2004.