De Novo Protein Sequencing from Tandem Mass Spectra

Tony L. Eng

Artificial Intelligence Laboratory
Massachusetts Institue Of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

The Problem: To find the primary sequence of a novel, previously unsequenced peptide from knowledge of its mass and access to its tandem mass spectrum.

Motivation: Proteins are essential to life, playing key roles in all biological processes: from enzymes that catalyze reactions, to antibodies in an immune response, from messengers in signalling pathways that allow a cell to react to stimuli, to secreted messengers that effect extracellular changes, and much more. Such is the extent of protein functionality to the survival of any organism. One of the first steps in understanding a protein is discovering its primary sequence.

The Edman Degradation reaction, an automated chemical process, is often used for the de novo sequencing of novel proteins, but it takes 30-60 minutes for each residue and there are other complications [1, 2]. Researchers have considered using mass spectrometry as a faster alternative.

Previous Work: Early de novo sequencing from mass spectra was performed by hand. Manual sequencing however is a tedious process that quickly gets complicated with complex spectra. A few computer algorithms for de novo sequencing of tandem mass spectra exist. These range from exhaustive sequence searching to stepwise generation of selective sequences, but of them are widely used [3], perhaps due to interest in more interactive forms of analysis [4], and/or low confidence in the predicted answers. So de novo protein sequencing is still largely an open problem [5, 3].

Approach: To overcome some of the obstacles (for example, the problem of gaps, underrepresented peaks in the spectrum, etc.) that the previous works encounter, we investigate a more global strategy simulated-annealing approach.

Impact: Rapid peptide sequencing of newly identified proteins.

Future Work: We will be analyzing this approach to evaluate its feasibility, performance and limitations.

Research Support: Early support was provided by an MIT Whitehead Training Grant in the Genomic Sciences, NIH/NHCRI HG00039, sponsored by Professor Paul Matsudaira, with supplemental support from a graduate student fellowship from the Program in Mathematics and Molecular Biology through Professor Bonnie Berger. Recent support has been through teaching assistantships. Research is currently under the supervision of Professor Tomas Lozano-Perez.

References:

[1] J. Yates, III and P. Griffin and L Hood and J. Zhou. Computer Aided Interpretation of Low Energy MS/MS Mass Spectra of Peptides. *Techniques in Protein Chemistry II*, 1991,477–485.

[2] W. Hines and A. Falick and A. Burlingame and B. Gibson, B. Pattern-Based Algorithm for Peptide Sequencing from Tandem High Energy Collision-Induced Dissociation Mass Spectra. *J Am Soc Mass Spectrom*, 1992, 3,326–336.

[3] M. Mann. personal email communication. April, 1998.

[4] C. Bartels personal email communication. April, 1998.

[5] V. Dancik and T. Addona and K. Clauser and J. Vath and P. Pevzner. De Novo Peptide Sequencing via Tandem Mass Spectrometry: A Graph-Theoretical Approach. *RECOMB*, 1999, 135–144.