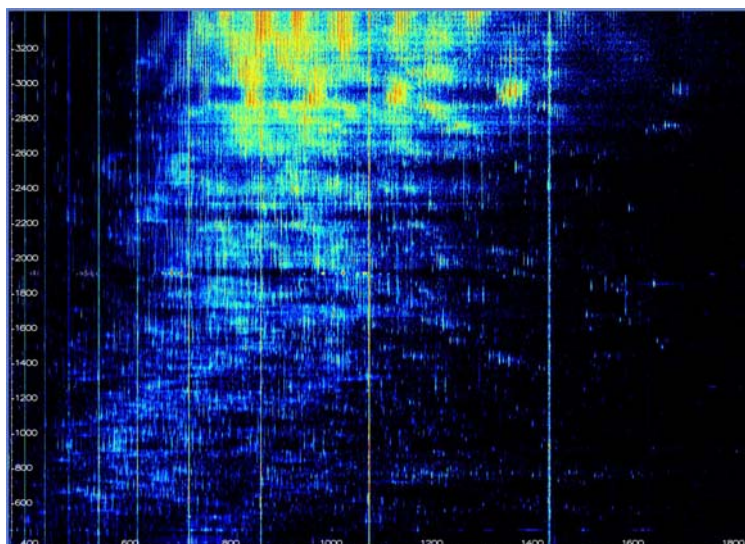


Development of an Expectation-Maximization framework for modeling protein identities and quantities in LCMS Experiments.

Introduction:

Proteomics methods attempt to accurately and comprehensively identify and quantify the protein constituents of complex mixtures. Comparisons across samples have been used to identify cellular functions and pathways affected by perturbations and disease, identify new components and changes in the composition of protein complexes and organelles and have led to the detection of putative disease



biomarkers. A significant challenge in mass-spectrometry based proteomics is the interpretation of the complex data generated by each experiment. In this approximately 3,000 megapixel image, what peaks are derived from peptides, and which are likely noise (either chemical or mechanical)? Having identified which signals are likely peptides, it then becomes necessary to characterize which peptides gave rise to which signals and how abundant those peptides were in the mixture. To date, naïve feature extraction methods have been successful in detecting obvious signals within the data, but have struggled to detect low lying signals and have struggled even more to explain the detected signals in terms of peptide identities and abundances.

Project Overview

We can re-interpret the challenge of explaining LCMS data as an incomplete data problem; under that guise an Expectation-Maximization framework may provide an appropriate solution. Such a solution will rely heavily on the development of a probabilistic model that maps from lists of peptides and their abundances to LCMS signals and vice versa. The significant experience our group has developed with modeling individual peptides signals should provide a strong basis for complex model exploration. For example, some signals in the data can be readily explained by peptides identified by MS/MS data. Given this set, it should become possible to generate a forward model of the signals expected to be generated from these peptides and some estimate of their abundances. By determining what portion of the data is explained by signals related only to the peptides identified by MS/MS it will then be possible to

determine what signal remain un-attributed. In an iterative step, one can attempt a variety of means to explain the remaining signals – for instance we can expand our model to attempt to explain remaining signals as alternate charge states of previously identified peptides, as degradation products of identified peptides, or as alternate peptides from proteins we suspect exist in the sample.

Deliverables

The following deliverables are expected at the end of the program:

1. A written report on the project including detailed description of the method developed for solving the problem and performance comparison with pre-existing methods such as matched filtering
2. Code implementing the method and data resulting from the project
3. Documentation for the code
4. A presentation of results and future directions to the SFCAP team