

Spielberg Family Center for Applied Proteomics Summer Project Description

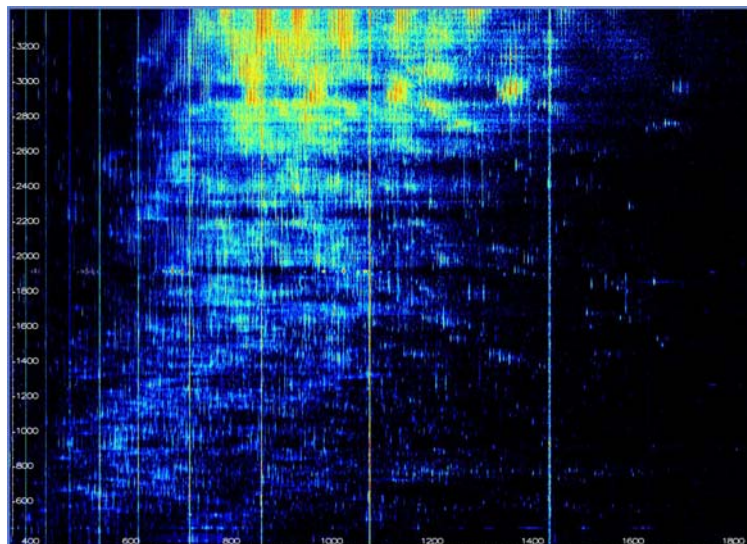
Project Title: Development of a probabilistic framework for estimating protein quantities in LCMS Experiments.

Introduction:

Proteomics methods attempt to accurately and comprehensively identify and quantify the protein constituents of complex mixtures, such as cell lysates and human plasma. 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.

Shown right is an example of a proteomics dataset - an approximately 3,000 megapixel image. Sophisticated tools have been developed to identify which peaks are biologic in origin. In addition, it is possible to assign a specific peptide sequence to many of these peaks to map them across experiments. However, it is not yet possible to accurately estimate the abundance of a peptide. Several factors confound this estimation including noise and signal suppression.



Project Overview

We can rephrase the challenge of explaining LCMS data as an incomplete data problem. Given an input set of spectra, approximate coordinates of peptide features and their identities one would like to return a matrix of abundances of these peptides across a range of experiments. Ideally this matrix of abundances will readily correlate to true abundances. Solving this problem will rely heavily on the development of a probabilistic model that maps from observed LCMS spectral intensities to abundances 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 should have correlated abundance as they are derived from the same protein and therefore the majority of fluctuation in their abundance should be due to non-biologic effects. Given an initial set of intensities, 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 the model explains it will then be possible to determine what signal remains un-attributed. In an iterative step, one can attempt a variety of approaches to explain the remaining signals - for instance we can expand our model to attempt to explain remaining signals as noise, signal suppression or peptides not in our model.

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