172. Algorithms Enabling Metaproteomics

Seungjin Na[1], Grant Fujimoto[2], Sangtae Kim[2], Brian LaMarche[2], Nuno Bandeira[1], Stephen Callister[2], **Samuel H Payne[2]*** (samuel.payne@pnnl.gov)

1 University of California, San Diego; 2 Pacific Northwest National Laboratory

http://omics.pnl.gov

Project Goals: This project is focused on improving algorithms for mass spectrometry data analysis of metaproteomics data. Recent advances in mass spectrometry and biological separations have dramatically increased the depth of proteomic discovery. Unfortunately, traditional computational workflows are in many cases preventing researchers from realizing these benefits for microbial communities. We propose to create a new generation of computational workflows to overcome the sensitivity limitations inherent in status quo data processing schemes.

Unlike laboratory experiments that focus on one single organism, biofuel production is often explored in experiments with either natural or synthetic communities composed of numerous organisms. To understand the active state of these communities it is advantageous to assay the proteome as opposed to the genome. Unfortunately current algorithms for peptide/spectrum matching are reliant on protein databases. The practical implication of this dependence for proteomics of communities (metaproteomics) is that the algorithms often fail to identify sufficient number of peptides and proteins.

The major goal of this project is to improve peptide and protein identification in community proteomics datasets. Thus, our algorithms must be designed to operate with millions of protein sequences (e.g. for an experiment with very deep metagenomic sequencing) or without sequences at all (for a project without matched metagenomics). To this end, we have designed several new algorithms: SpectralNetworks with AlignGF and tag-filtering, Informed Quantitation of Top-down proteomics, and alignment via spectrum/spectrum matches.

SpectralNetworks has been effectively used to discover protein post-translational modification and mutations. Although it proved to be useful in analyzing small spectral datasets, it poses challenges in its reliability and scalability for large spectral datasets. The key problem in SpectralNetworks is distinguishing between correct and incorrect spectral pairs. AlignGF (Alignment generating function) can be used to calculate theoretical distributions of scores for all possible alignments between two spectra, and provide a rigorous solution to the problem of computing statistical significance of spectral alignments. A binomial distribution is assumed to regulate the probability of aligned (or matching) peaks among total peaks of a spectrum. The implementation of AlignGF has dramatically improved the sensitivity of SpectralNetworks in large datasets, achieving ~75% sensitivity at ~90% precision for spectral pairs of mutated peptides. A second improvement to the SpectralNetworks algorithm is the inclusion of tag-based filtering. The previous algorithm needed to compute all pairwise spectral alignments for the construction of spectral networks, requiring significant computational resource.

Properly aligned spectra, from peptides with similar sequences, share a substring of the sequence. Therefore, we use spectrum tag generation to filter which spectral pairs can be aligned. The tag-based filtering currently accelerates SpectralNetwork generation by 200x with little loss in sensitivity. As mass spectrometers continue to improve, the analysis of intact proteins is becoming more common. Top-down proteomics is advantageous as it can fully characterize a protein with all its post-translational modifications which are critical for function. We are developing algorithms for top-down, in anticipation of their use in community studies of biofuels, particularly for simplified or synthetic communities. Our

current algorithmic progress is with a quantitation tool called IQ, which accurately measures the abundance of each individual proteoform in a sample.

As a final algorithm for metaproteomics, we are developing methods to align mass spectrometry datasets without the requirement of peptide identifications. LC-MS alignment is crucial when researchers are looking for common features across several LC-MS/MS runs. Instrumental variation, especially in the chromatographic dimension can hinder the ability to find the same species. When genomic information is incomplete or missing (e.g. metaproteomic applications) existing alignment techniques that require peptide identifications for alignment may fail. Using MS/MS spectra to define these anchor points can provide a highly confident alignment, even when there are few anchor points as might exist in rapidly changing communities.