

221. Pan-omics: Bioinformatics Developments

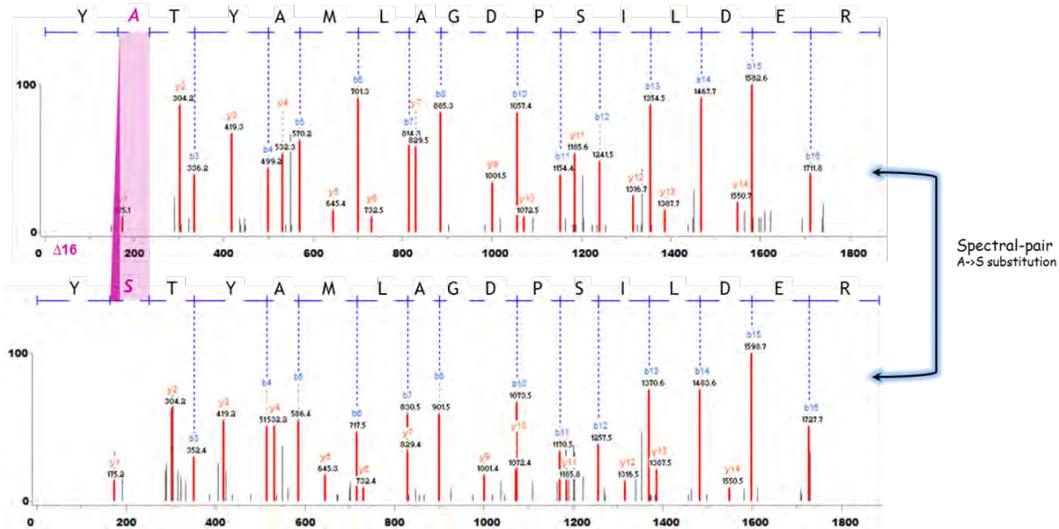
Samuel H Payne, Jason E McDermott, Nick Cramer, Joshua N Adkins* (joshua.adkins@pnnl.gov), Mary S Lipton, Richard D Smith

Biological Sciences Division, Pacific Northwest National Laboratory, Richland WA

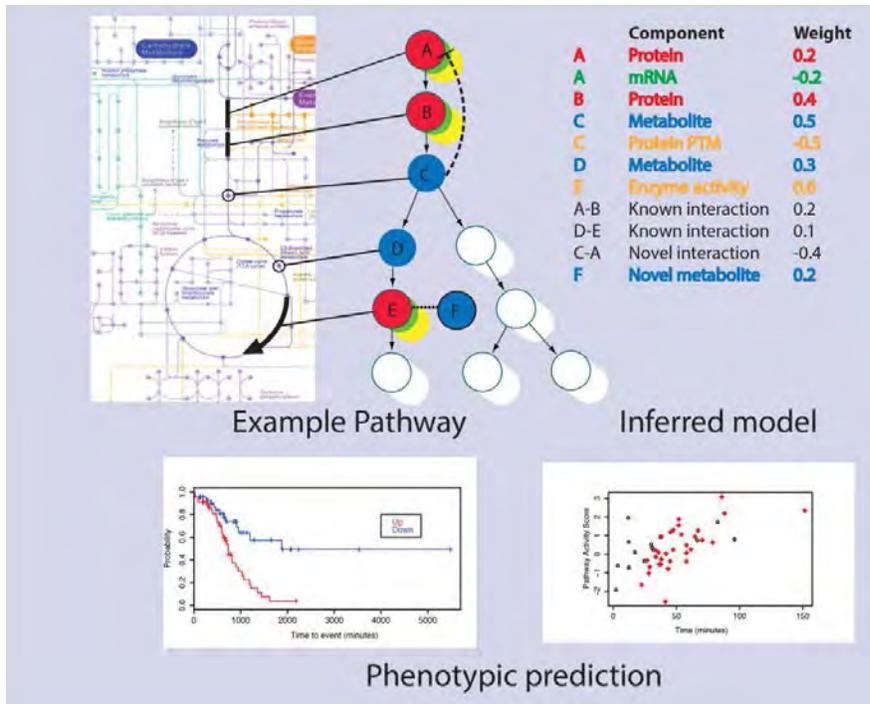
<http://omics.pnl.gov>

Project Goals: The Pan-omics Program scope includes creating rich multi-omic datasets for GSP relevant biological systems such as microbial communities and plants. The goal of the bioinformatics component within the Pan-omics Program is to build tools that can deal with increasing data throughput, and both speed and assist in data analyses and interpretation to help advance the knowledge about these environments/systems. We have recently focused on three critical areas: improved methods for proteomic characterization of microbial communities and other natural systems, integrating diverse pan-omic measurements, and advanced methods for collaborative and visual data analysis. These three efforts follow the natural flow of data as it is processed, analyzed and interpreted for biological conclusions.

For mass spectrometry-based metaproteomic measurements where the corresponding metagenome information is unavailable, or where strain heterogeneity prevents complete resolution of the metagenome, traditional proteomics data analysis algorithms are insufficient. We are developing a new method, called Library Network Alignment that utilizes previously annotated spectra in a library to help identify the spectra from the metaproteomics sample. The core of this algorithm is the ability to recognize similar spectra from similar peptides and translate the sequence identification. The figure below is an example of spectral alignment between peptides with an S->A mutation.



After quantifying proteins, metabolites, mRNA, etc. researchers need a way to coherently identify which biomolecules are correlated to their experimental inquiry. LEAP is a lasso-based statistical regression tool that identifies components of networks and pathways that are predictive of phenotype and experimental variables. LEAP integrates all pan-omic measurements as well as other high-throughput technologies into a single model, thus leveraging the unique information of each data-type.



Analysis of large and heterogeneous data requires multiple domain experts, and is a common pan-omics bottleneck. Moreover, the biological conclusions which are derived from data are most often discovered by non-computational scientists. Therefore, data analysis need to be both collaborative and visually accessible. We are developing the Active Data Canvas as a visual analytics tool that greatly facilitates collaboration and data browsing by providing on-the-fly statistical analyses, integrated and interactive data visualization, and a sharable thought-space for annotating significant data or conclusions.

