The PNNL Proteomics Biodiversity Library

Samuel H Payne, Matthew E Monroe, Gary R Kiebel, Grant Fujimoto, Christopher Overall, Michael Degan, Bryson Gibbons, Samuel O Purvine, Mary S Lipton, Richard D Smith
Biological Sciences Division, Pacific Northwest National Laboratory, Richland WA

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle *Since 1965*

Overview

- Announce publication and deposition in open repositories of extensive proteomics data for 112 microbial species
- Proteome coverage averages 80% of functionally annotated genes in KEGG pathways
- Design framework for expanding the Biodiversity Library with submissions from the community
- Outline algorithms for using the Library in metaproteomics data analysis

Introduction

- A diverse set of organisms has been run by a variety of groups around the world, but they are not analyzed or processed in a standard way.
- Large data stores are not organized and colated for easy public consumption.
- Data from a phylogenetically diverse set of organisms would be a benefit to bioinformatics algorithm development.
- Publically available needs to be presented in a variety of ways for different uses by various communities.

Methods

- Data in the Library from pure cell culture experiments
- Data represents 10-15 years of collection
- Variety of sample prep methods, most often whole cell lysate and protein extraction
- MS/MS data acquisition on Thermo LTQ, Orbitrap, Velos instruments in DDA mode
- MS/MS data analysis via MSGF+
  - RefSeq protein annotations
  - PTMs: methionine oxidation
  - Quality filtering, q < 0.0001
  - 2 peptides/protein

Publically Available Data

- Global proteomics data for 112 bacteria and archaea
  - Raw mass spectra (vendor and open formats)
  - Peptide identifications from MSGF+ (mzid)
  - Condensed spectral library (bibliospec)
  - Proteins mapped to KEGG pathways
  - On average, 2100 proteins per organism
- Data uploaded to ProteomeXchange via MassIVE
- MSV000079053
- PXD001860
- 13 TB of data

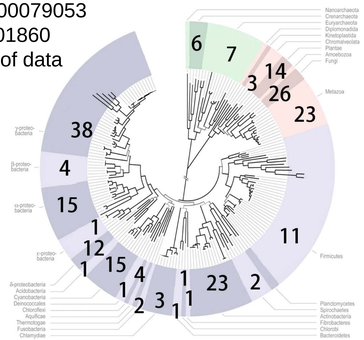

Figure 1 – Diversity. Organisms in the Library are overlaid onto the Tree of Life, showing the breadth of coverage for any particular taxa.

Results

Functional Coverage

- Using standard RefSeq accessions, all protein identifications are mapped onto KEGG functional classifications, including pathways
- For any specific organism and Pathway, the Library shows which proteins are identified
- For proteins with a known function, coverage with MS/MS data is generally high.

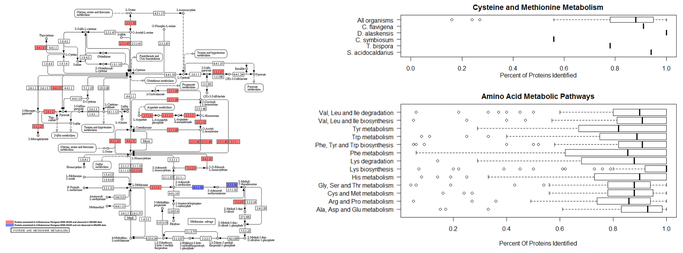

Figure 2 – Protein and Pathway Coverage. For a given pathway and organism, the Library exposes how many proteins in the pathway are observed in MS/MS data. On the left we see the Cysteine and Methionine metabolism pathway for C. flavigena, where nearly every protein was observed. When looking at all organisms, one can identify the general coverage of a given pathway (right).

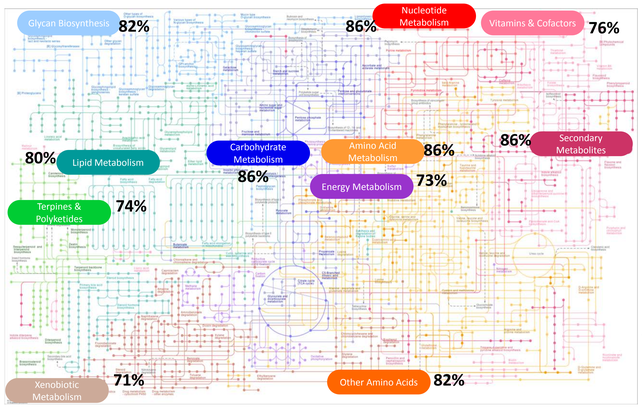

Figure 3 – Average Coverage of Functionally Annotated Proteins. KEGG organizes related pathways into superpathway groups, e.g. the 13 pathways for amino acid metabolism. The library identifies the average coverage of a pathway across all organisms (Fig 3). When viewed on a global scale, the average coverage of any pathway in a super pathway group is also exceptionally high, e.g. 86% for any amino acid metabolic pathway in any organism.

Related Peptides

- Numerous sequence similar peptides exist in the library, originating from aligned regions of homologous proteins
- Related peptides from different organisms offer new perspective for machine learning projects

Figure 4 – Homologous Peptides. Homologous proteins from different organisms typically have peptides observed in MS/MS data (left). Similar peptides from such alignable regions of a protein typically have very similar fragmentation patterns (right).

Community Re-use Projects

- We encourage the community to download the library and explore: fragmentation fundamentals, proteotypic peptides, LC elution prediction, etc.

Conclusions

- Promote free and open science through the public release of >13 TB of proteomic mass spectrometry data
- Enable spectrum library searches on any model organism

Contribute Data

- We welcome new organisms or conditions that significantly add to the breadth and/or coverage of the Library
- PSM identification, stringent quality control (q < 0.0001)
- Data Deposition
  - ProteomeXchange consortium (e.g. MassIVE)
  - Tag submission with "Biodiversity Library"
  - Send the Bibliospec Library to PNNL

Career Opportunities: For potential openings in the Integrative Omics Group at PNNL please visit http://omics.pnl.gov/careers

CONTACT: Samuel H. Payne, Ph.D.
Biological Sciences Division
Pacific Northwest National Laboratory
E-mail: samuel.payne@pnnl.gov

Acknowledgements

This work was funded by an Early Career Award from the U.S. Department of Energy (DOE) Office of Biological and Environmental Research (OBER) to SHP. Samples were analyzed under the Pan-omics program supported by DOE/OBER using capabilities developed under the support of NIH National Institute of General Medical Sciences (GM103493), and the Pan-omics program and performed in the Environmental Molecular Sciences Laboratory, a DOE OBER national scientific user facility on the PNNL campus. PNNL is a multiprogram national laboratory operated by Battelle for the DOE under contract DE-AC05-76RL01830.

www.omics.pnl.gov