DATA MANAGEMENT PLAN

This proposal will generate three types of proteomics data; de novo peptide sequences, metaproteomic identifications, and targeted peptidoglycan quantifications. All data will be archived in accordance with NSF policy (see AAG Chapter VI.D.4) on the dissemination and sharing of research results. We will submit all raw data to PRIDE and all processed data to the Biological and Chemical Oceanography Data Management Office (BCO-DMO; http://bcodmo.org/resources) for curating. Experimental data will be published in a timely fashion.

Proteomics analysis generates large volumes of data (~1 gb per run). Access to and manipulation of the data requires large storage and computational capacity. The MMRC facility that houses the mass specs has an open access archive of data and adequate computational space for manipulating data. We use the University of Washington's Hyak Cluster Computer Facility (http://escience.washington.edu/content/hyak-0). File format standards for these data sets and electronic dissemination and preservation plans are as follows: The raw data and instrument settings from each mass-spec run are saved in native file formats. For proteomics analysis, the files will be converted to the open community standard mzXML file format. These data will be accessible upon request through the collaboration file system. Metadata and tabular results will be made available through the collaboration file system in a timely fashion and deposited with federally funded clearing houses for electronic dissemination.

The compute node for data workup is integrated into the Hyak scalable cluster compute facility. As of June 2014 Hyak comprises >1000 compute nodes. Hyak users are guaranteed immediate access to their own nodes when launching jobs. They also have access to idle nodes throughout the cluster by way of a backfill queue. The MMRC operates one computer node within the Hyak cluster.

Archival Data Storage: Lolo is a file-based storage service for research computing customers at UW. It includes two file systems, Archive and Collaboration. The archive service is intended as a repository for data that you may rarely access but that you want to ensure is safe and available over the long term. Users write files to disk as they would with any other network file system. Within a day all files are automatically transferred from disk to tape in a primary campus datacenter. A second tape copy is created in another campus datacenter within an additional day. Recall of files is automatic and accomplished by simply opening and reading the file. The MMRC currently provides users with up to 8 TB of archival data storage.

We retain rights to "first publication" of our data. This data management plan was written consulting NSF document 04-004 "Division of Ocean Sciences Data and Sample Policy".