Computational Approaches for Aggregating and Scoring Protein-protein Interaction Data

William R. Cannon*[1](william.cannon@pnl.gov), W. Hayes McDonald[2], Don S. Daly[1], Denise Schmoyer[2], Gregory B. Hurst[2], Manesh B. Shah[2], Brian S. Hooker[1], Vladimir Kery[1], Stephen J. Kennel[2], H. Steven Wiley[1], Michelle V. Buchanan[2]

[1]Pacific Northwest National Laboratory, Richland, WA, and [2]Oak Ridge National Laboratory, Oak Ridge, TN

The GTL CMCS informatics pipeline for *Shewanella oneidensis* MR-1 and *Rhodopseudomonas palustris* protein-protein interaction networks melds information from two sources: data collected internally, including both high-throughput pipeline (see companion abstract: "Comparison of conserved protein complexes across multiple microbial species to evaluate high-throughput approaches for mapping the microbial interactome") and more targeted validation (see companion abstract: "Complementary Assays for Validating Protein Interactions Identified by High-throughput Screening Techniques") assays, and external data sources that both identify interactors and add biological context. A key function of this integrated pipeline is to assess the quality of new data and the processing pipeline itself using: targeted experiments, QC standards, technical and biological replicates, statistical modeling, and statistical process control. Protein interaction networks are then inferred using two different methods. A Bayesian method gives a global analysis of the entirety of the data including false positive and false negative error rates, while a binomial-based maximum likelihood method gives information on error rates of individual bait proteins. This information is captured and published on a web interface (see companion abstract: "The Microbial Interactome Database: An Online System for Identifying Interactions Between Proteins of Microbial Species"). In addition, new decision tools are being developed to automate target selection with the goal of picking the protein targets that will best increase the resolution and breadth of the protein interaction networks.