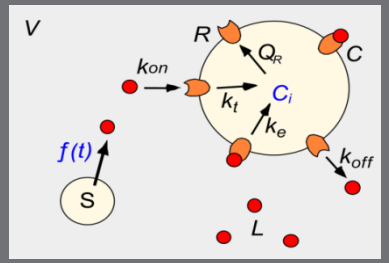
Computational Systems Biology

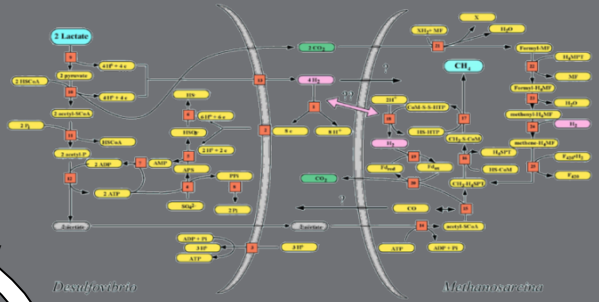
- Enabling fundamental advances in the predictive modeling of biological networks and biomolecular systems through the application of high performance computing

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Understanding systems biology with computing, analysis and visualization



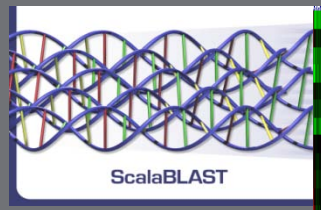
Metabolic and regulatory models

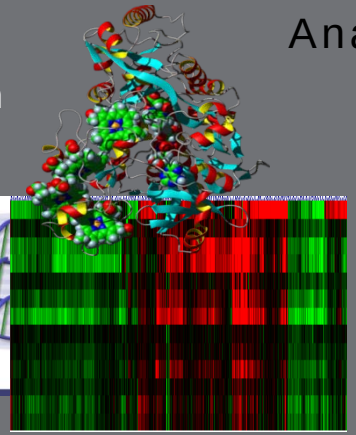


Theory

Analysis Experiment

Advanced Computation and Information Management





Advanced High-Throughput Technologies

Pacific Northwest NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

Systems Biology

- PNNL's effort in systems biology research includes
  - Large-scale modeling and simulation—from molecules to communities
  - Management and integration of large-scale data from
    - Instruments, simulations, sensors, and calculations
  - Algorithms for data analysis and visualization
    - Network analysis, visualization, inference
    - Interactive hypothesis discovery and evaluation
    - High-performance task scheduling

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Leadership in Systems Biology

- PNNL is leading the way in bringing HPC to bear on challenges in systems biology with
  - High-impact science articles
  - Highly used/cited software
  - Thought leaders in integration and visualization
  - External bio projects on cluster
- World-Class High-throughput Proteome Measurement Capabilities

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Systems Biology – Integrated, multidisciplinary, multi-institutional solutions

- PNNL's unique combination of
  - Unclassified large-scale computing resources
  - Visualization expertise
  - Scientific HPC software development
  - Data management and integration technologies
  - Scientific collaborations with academic, industrial and government partners

Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by Battelle *Since 1965*

Proteomics – Advancing human health, environmental, and energy research

- Extensive investments in advancing proteome measurement platforms and the supporting informatics infrastructure
- Assembly of a large and expert team with diverse skill sets needed for high-throughput operations
- Operational framework for rapidly translating and efficiently implementing concurrent technological advances (including many originating at PNNL/EMSL)

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Bringing HPC to bear on challenges in systems biology

- Openly available, highly scalable and portable software
  - ScalaBLAST, Polygraph
- Workflow, data management, software
  - BRM, EDMS, MeDICi
- Basic and applied research at community and metagenome scale
  - Renewable energy, environmental cleanup
- Algorithm development for next-generation technologies

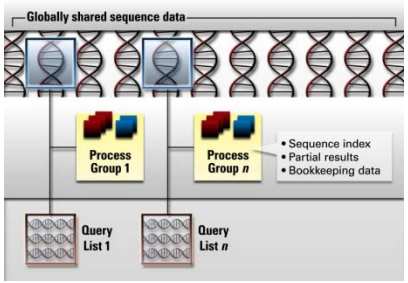
Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Fundamentals to high-throughput measurement capability

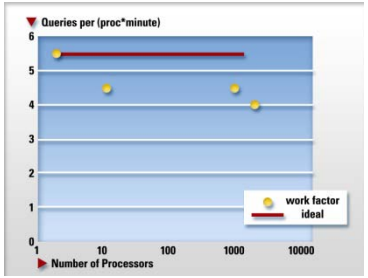
- Accurate mass and time (AMT) tag high-throughput proteome measurement strategy
- Extensive proteomics pipeline based on high resolution separations combined with high sensitivity MS and advanced informatics tools (http://omics.pnl.gov/),
- Infrastructure for processing and managing large volumes of data.
- Development of next generation high throughput measurement platforms.

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

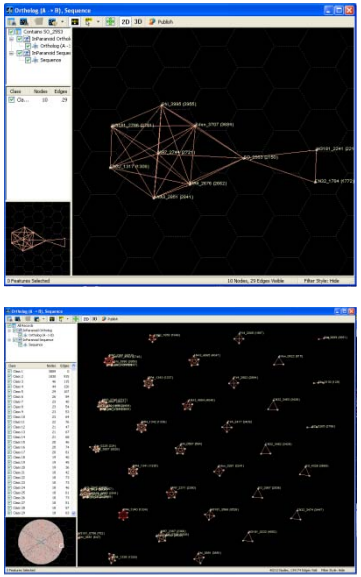
High Performance Applications: ScalaBLAST



Efficient parallelization of sequence alignment



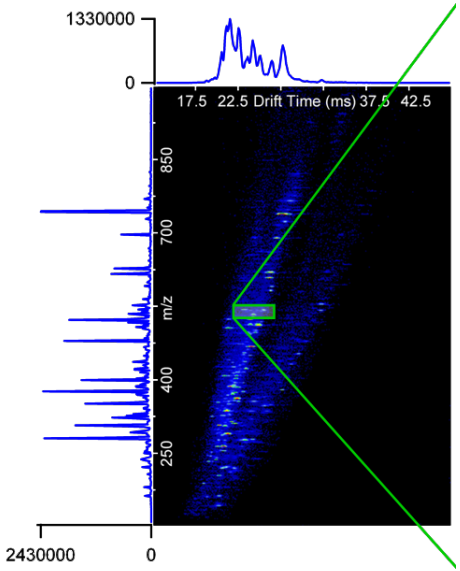
Rapid execution on large-scale systems

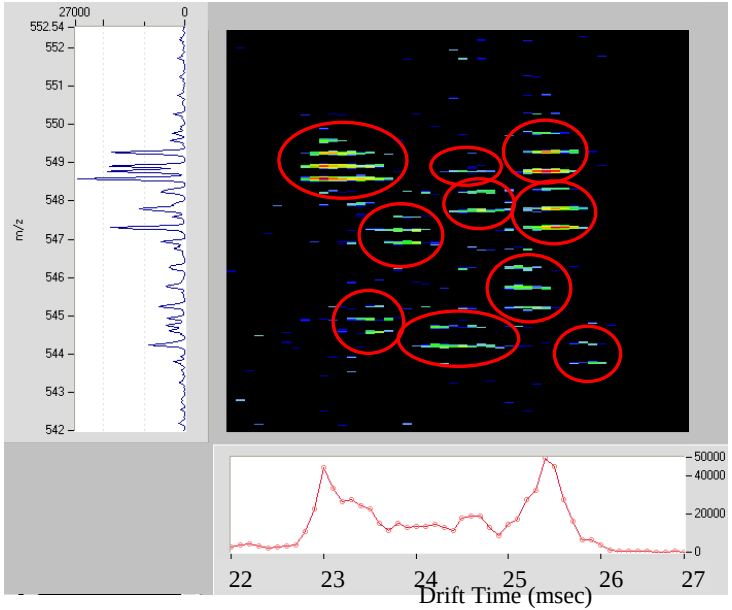


Discovering, evaluating hypotheses in real-time at *multiple genome scale*

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

High Performance Applications: Multidimensional feature extraction





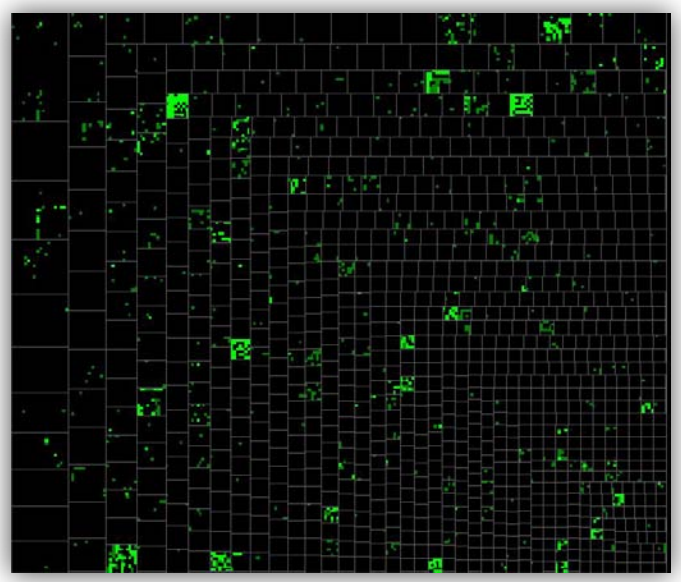
Detail of single time point showing clear separation using multiple dimensions

Pacific Northwest NATIONAL LABORATORY

Proudly Operated by Battelle *Since 1965*

High Performance Applications: Clustering

- Global Ocean Survey Metaproteomics analysis
- HPC resources used for:
  - ScalaBLAST clustering of sequence data.
  - Data analysis using peptide identification algorithms.
- Over 6 million entries in protein database

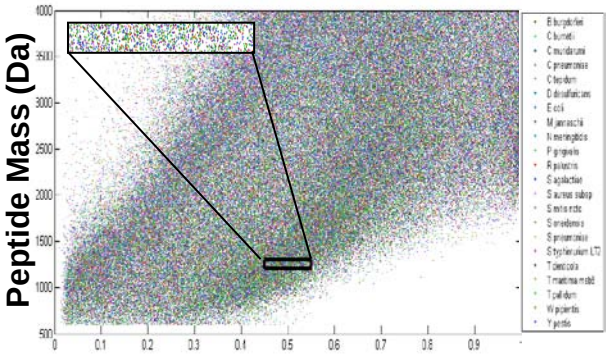


Example display of identified clusters of similar protein sequences

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

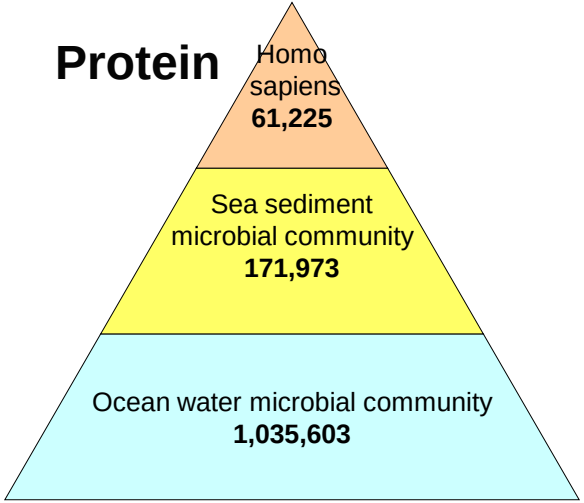
High Performance Applications: Mapping metagenome and community spectra to biomolecules

Peptide



- 2.8 million peptides from 25 microbes
- Dense central region has high ambiguity in identification, but precision in measurements improves confidence.
- *AD Norbeck et al., J. Microbiological Methods, 67(3) 2006, 473-86*

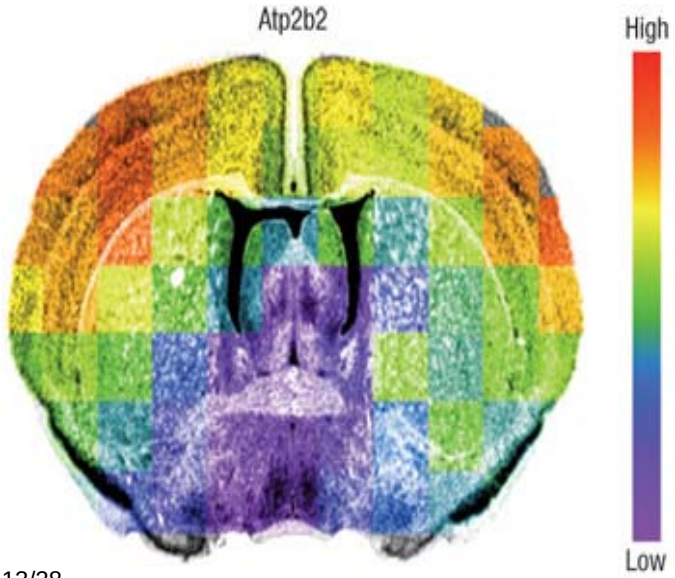
Protein

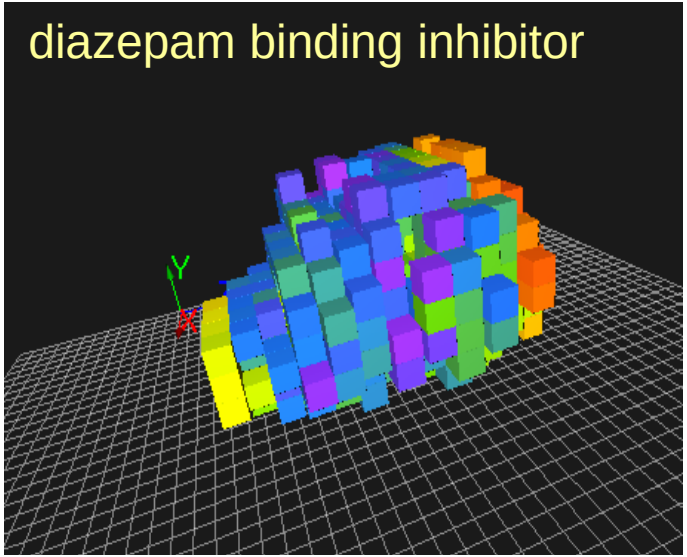


Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

High Performance Applications: Mapping biomolecules in tissues

- Mouse brain proteomics using analysis of ~700 tissue (voxel) samples and providing spatial distributions on more than 2000 proteins and quantitation for >1000 proteins

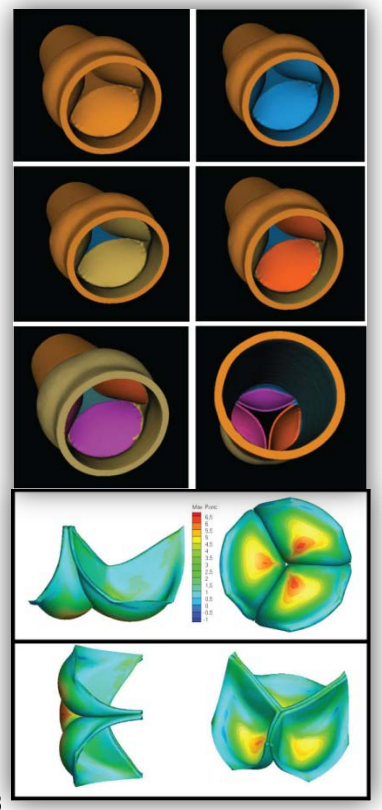




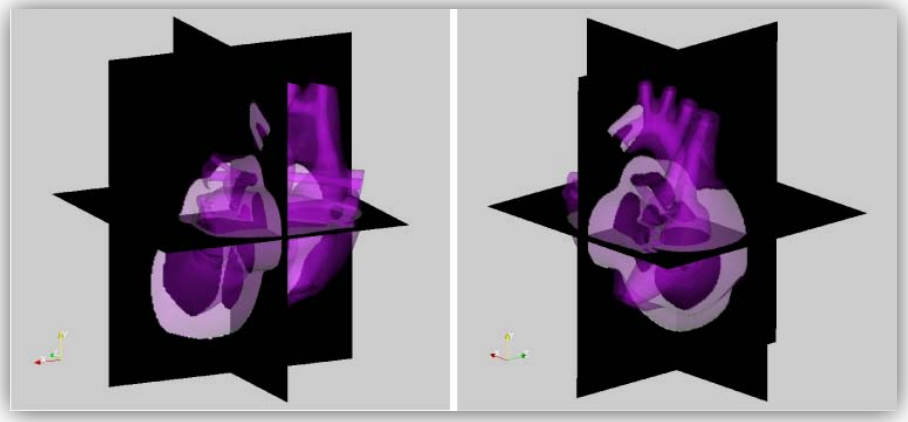
Collaboration with Desmond Smith (UCLA)

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

High Performance Applications: 3D Tissue modeling



Heart valve simulations include complex interactions between fluid and a non-stationary boundary—the heart tissue itself.



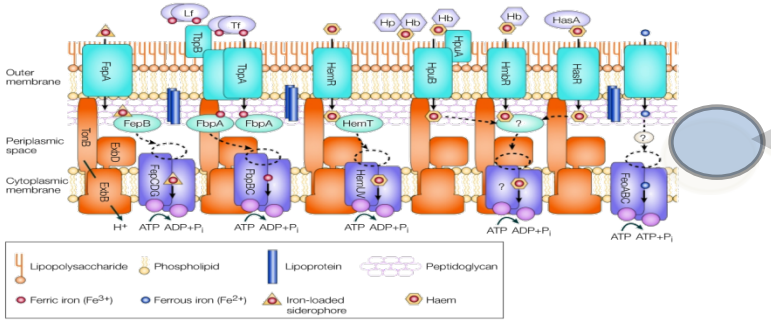
Modeling fluid flow in the heart involves complex dynamic interactions between the mechanical pump system and the fluid response to this pump.

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

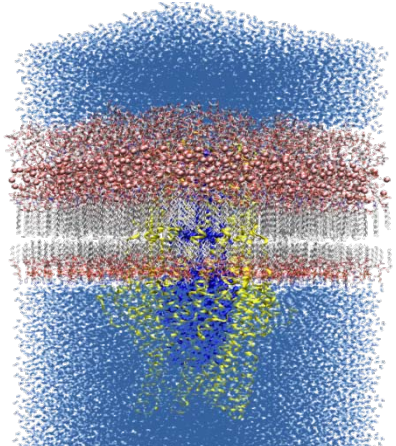
High Performance Applications: Massively parallel molecular dynamics

Atomistic simulations of bacterial outer membrane proteins of *Pseudomonas aeruginosa* are used to address critical health and environmental issues.

- Infections by P. aeruginosa are the major cause of morbidity and mortality of patients with cystic fibrosis, the most common inherited lethal disease in Caucasians (1 in 2,500 live births).
- The evolving problem of Pandrug-resistance in P. aeruginosa which has an extremely low antibiotic susceptibility due to the concerted action of multi-drug efflux pumps and the low permeability of the outer membrane.
- High performance modeling and simulations carried out at PNNL are providing insight into the mechanism of small molecule transport across the bacterial outer membrane.



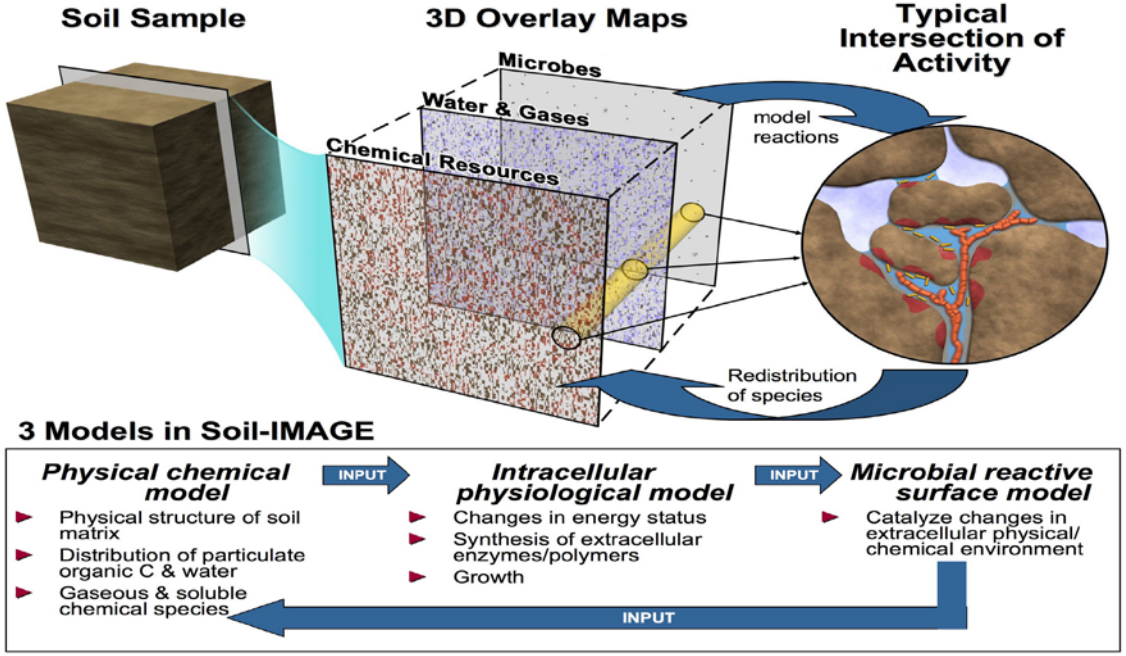
Modified from Faraldo-Gomez and Sansom, 2003



Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Integrative Modeling: Terrestrial microbial communities and their environment

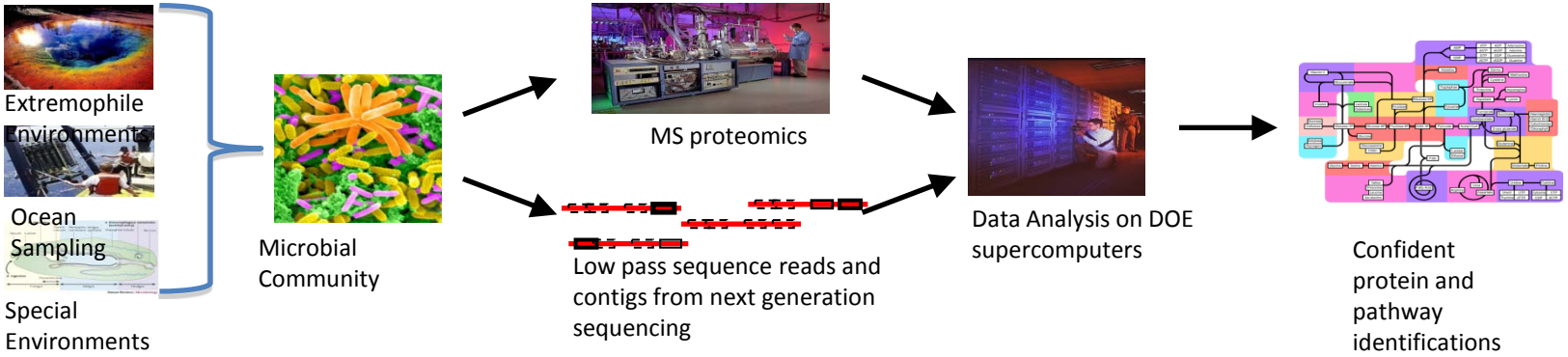
- Soil-IMAGE project integrates known physical, chemical, and biological properties of soil systems into a unified mathematical model.
- This model employs a discrete-continuum hybrid representation, and it is being used to make predictions about the microbial community dynamics in soil aggregates.



Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Integrative Modeling: Understanding the behavior of communities from molecular measurements

- The lack of reliable genome sequences currently limits the effectiveness of proteomics studies because of the difficulty in identifying peptides.
- The goal of this DOE BER/OASCR project is to develop novel methods needed to describe the proteins and metagenomic functional processes occurring within unsequenced microbial communities being investigated as part of DOE's missions in carbon sequestration, bioremediation and bioenergy research.

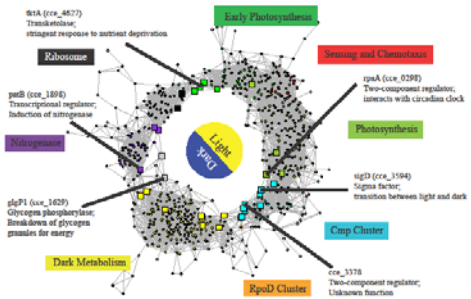


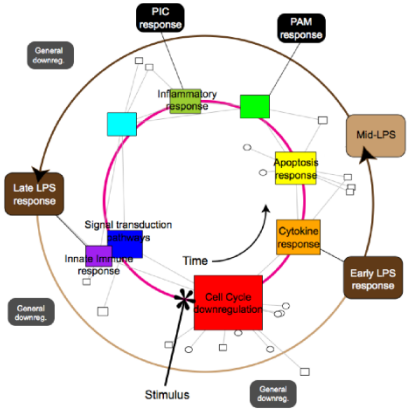
MSPolygraph: Identifying Proteins and Pathways from Microbial Communities

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Graph and Network Applications: Modular network modeling

- Networks are fundamental to many areas of biology
  - Gene regulatory interactions
  - Protein/protein interactions
  - Coexpression networks
  - Time-course expression
- Modular networks are a way to condense experimental data into high-level functional units.

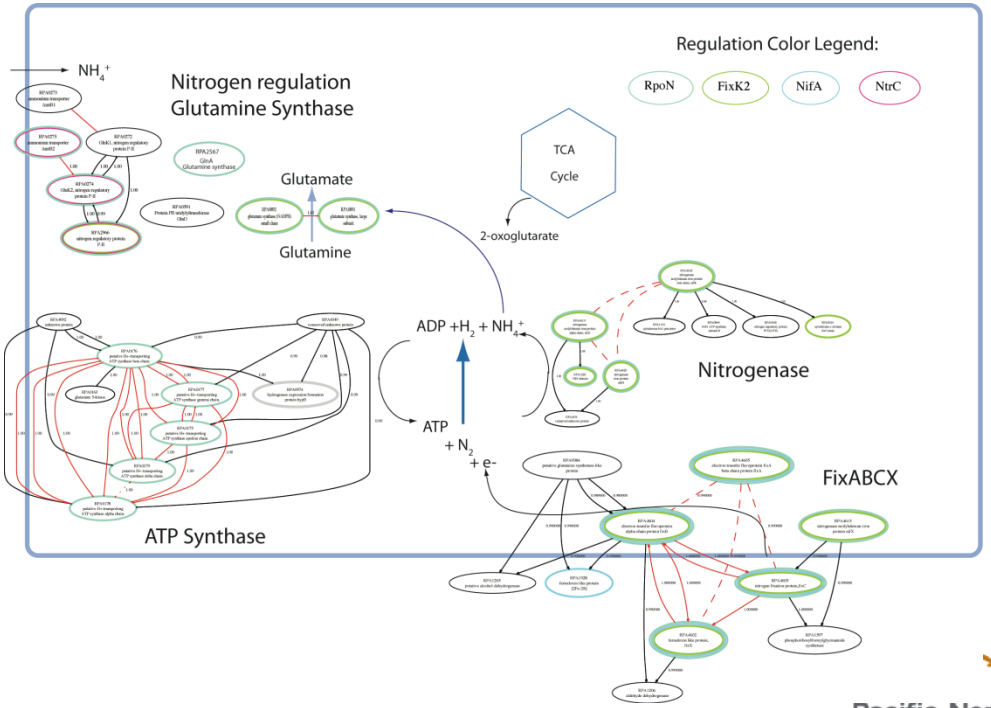




Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Graph and Network Applications: From data to proteins and pathways

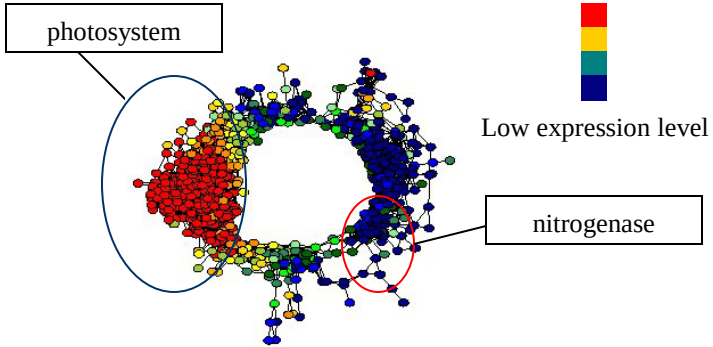
- **DOE GtL: Genomics:** Advanced mathematical models and computation leads to greater understanding of bioenergy pathways

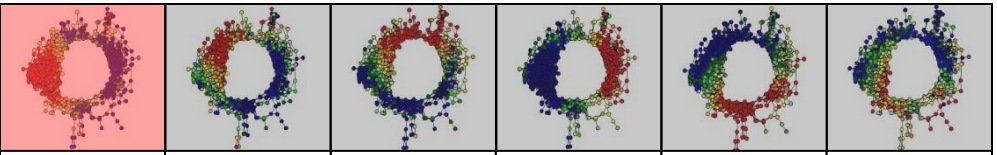


Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Graph and Network Applications: Understanding cyclic patterns in microbes

Relevance networks show 'flow' of gene transcription in *Cyanothece* during light/dark cycle.

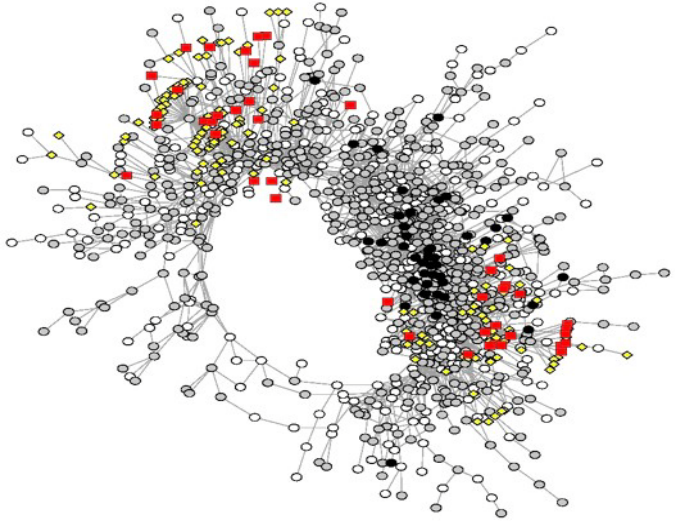




L1: photosystem genes are at high expression levels.	**L5:** Some photosystem genes still 'on'.	**L9:** Photosystem genes returning to baseline levels.	**D1:** Photosystem genes 'off'.	**D5:** Nitrogenase genes at high expression levels.	**D9:** Nitrogenase ramping down, photosystem starting again.

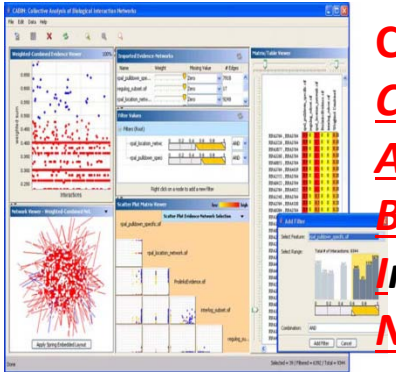
EMSL Grand Challenge in Membrane Biology Understanding energy production in blue-green bacteria during the light/dark cycle.

Integrated relevance networks show where metabolites (red, black nodes), proteins (clear nodes) appear together in the light/dark cycle.



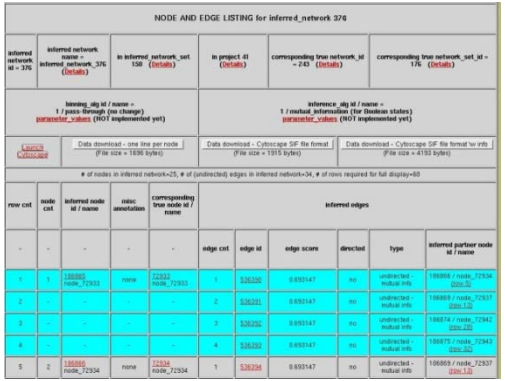
Graph and Network Applications: Cabin/Sebini for inferring network interactions

- SEBINI provides access to several network inference algorithms for use on gene expression or protein abundance data.
- CABIN allows simultaneous exploratory on-screen analysis of multiple inferred interaction networks.



CABIN:
Collective Analysis of Biological Interaction Networks

Use of the CABIN software to analyze experimental interactions for Rhodopseudomonas palustris using data obtained from 1082 bait-prey experiments conducted using affinity purification technique. The interactions of interest are selected in blue.



SEBINI:
Software Environment for BIological Network Inference

Node-edge table of transcriptional regulatory network inferred using Context Relatedness Likelihood (CLR) Algorithm in SEBINI.

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

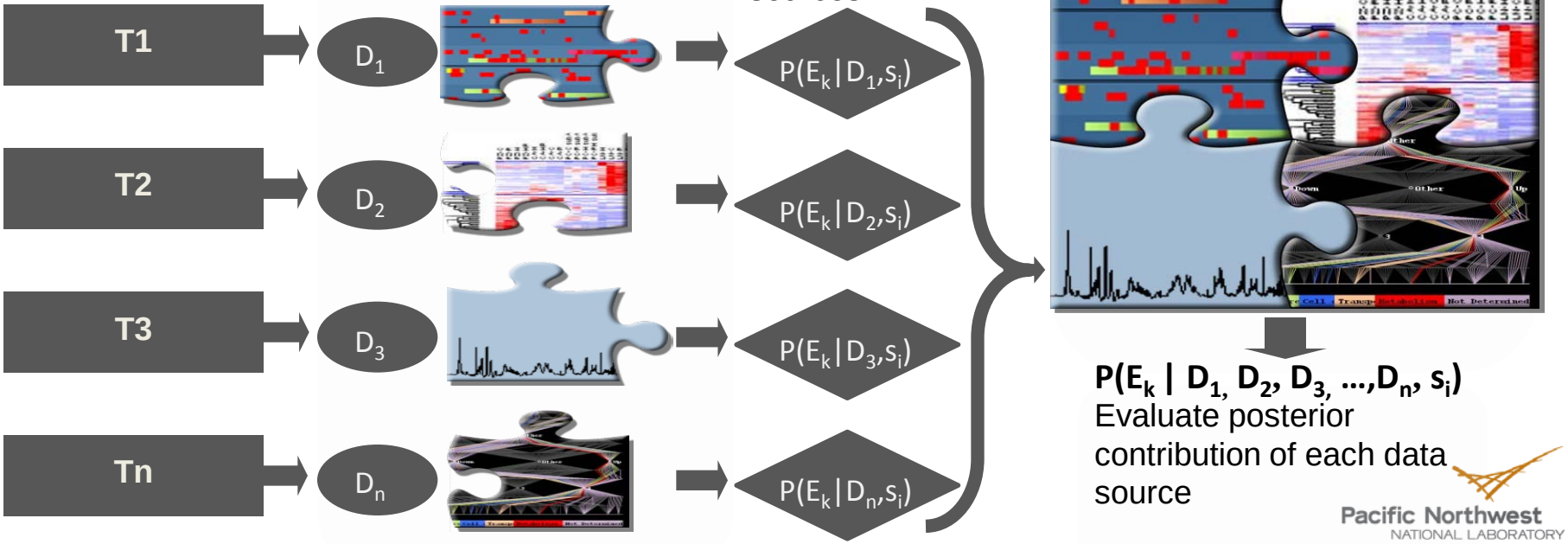
Statistical Methods: VIBE – Bayesian data integration

Technology

Map experimental data sources to common probabilistic form

Utilize probability mappings as diagnostic capability of individual data sources

Assume independence to integrate probabilities across data sets



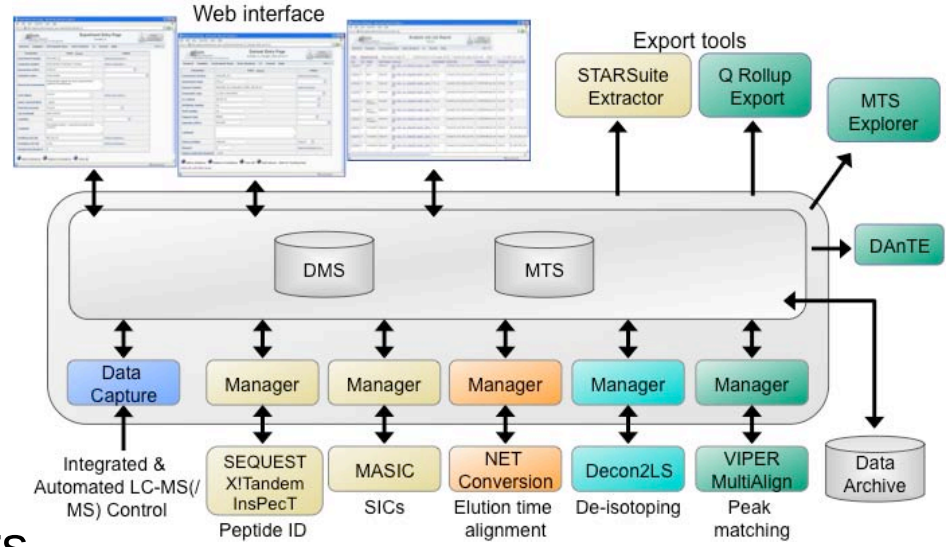
$P(E_k \mid D_1, D_2, D_3, \ldots, D_n, s_i)$

Evaluate posterior contribution of each data source

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Large Scale Data Management: Architectures for managing proteomics data

- Proteomics data management expandable architectures supporting:
  - 150+ tera bytes of data
  - Over a billion mass spectra
  - 200+ processors
  - Data dissemination
  - Data archival
  - Automated data capture
  - Web-based interface to users
  - Thousands of “omics” analysis per month

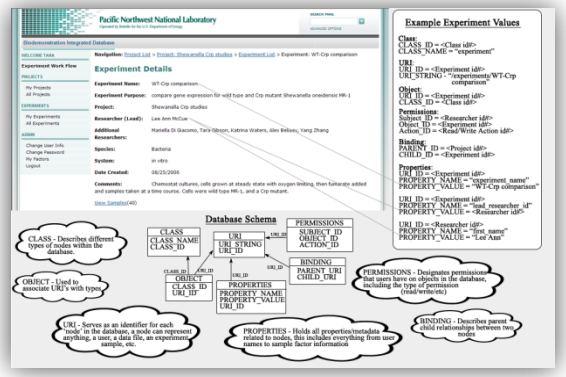


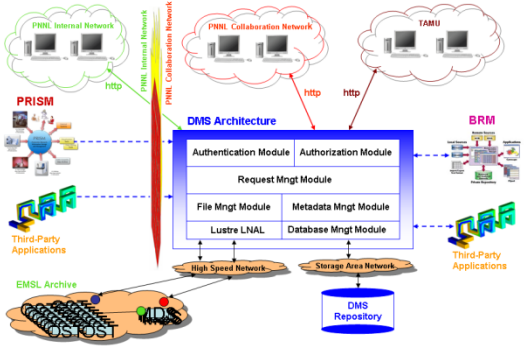
Pacific Northwest NATIONAL LABORATORY

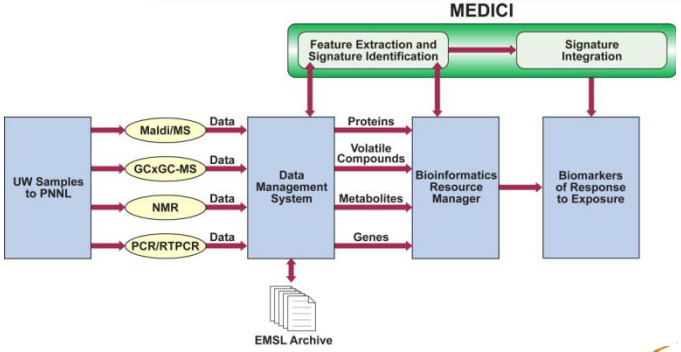
Proudly Operated by Battelle *Since 1965*

Large Scale Data Management: EDMS and EMSL archive

- EMSL archive is a large-scale storage resource for computational and experimental users
  - Instrument output
  - Simulation output
- EDMS enables
  - Rapid storage
  - Searching
  - Tool integration





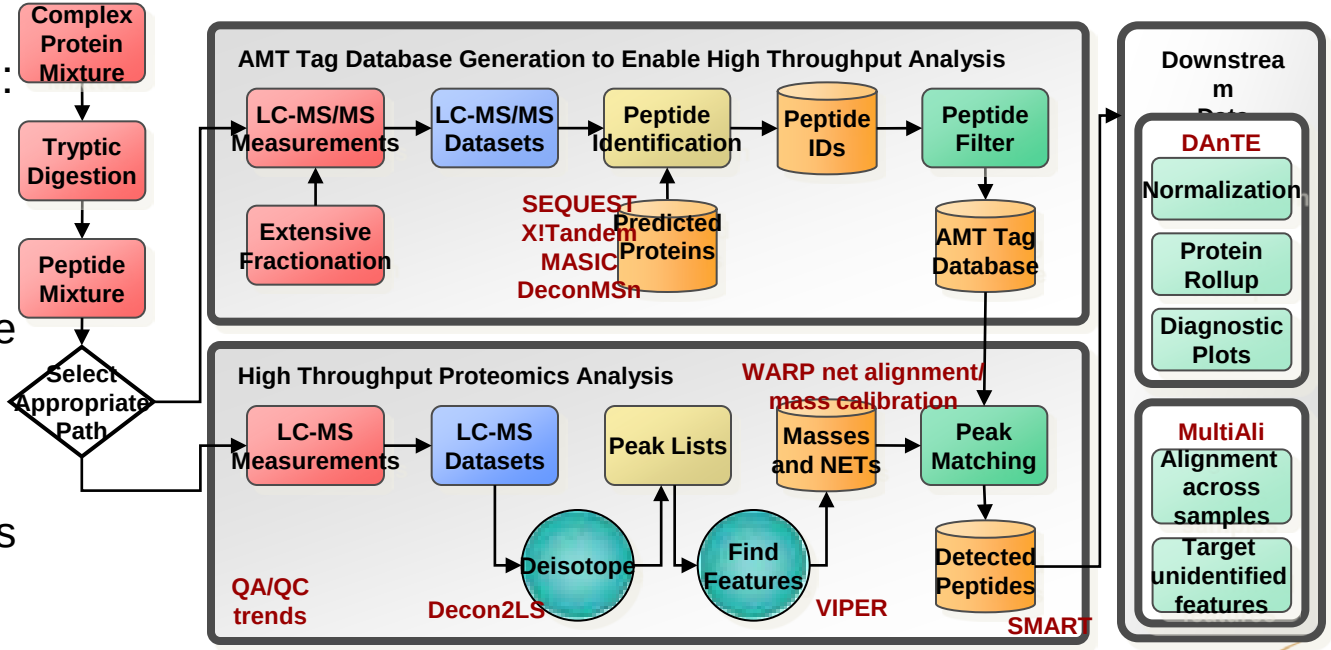


Pacific Northwest NATIONAL LABORATORY

Proudly Operated by Battelle *Since 1965*

High Throughput Data Analysis: Proteomics pipeline

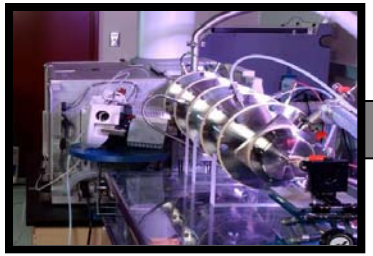
- Proteomic analysis pipeline automation:
  - Complex multi-path workflows
  - Analysis performed with tools from multiple sources
  - Heterogeneous data file formats
  - 1000s of data files contribute to one proteomics analysis



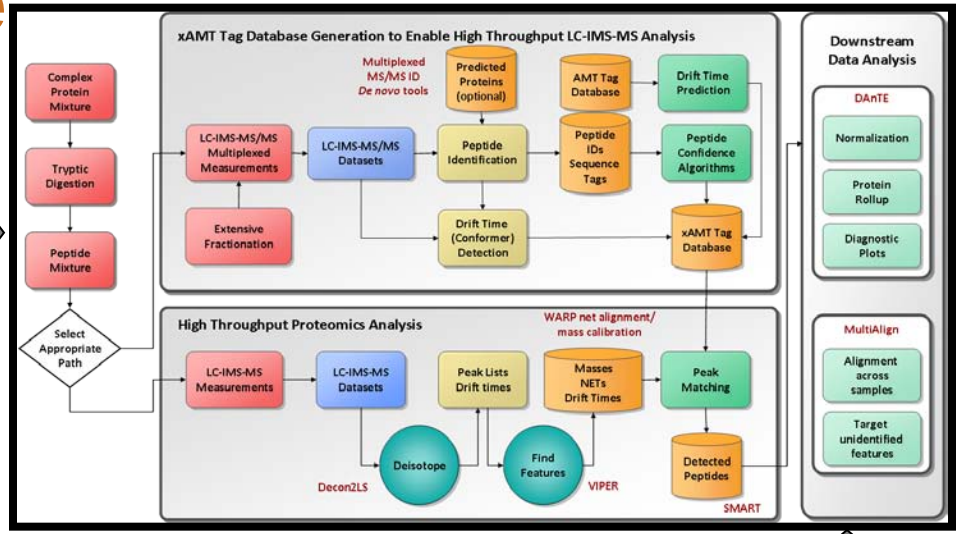
Pacific Northwest NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

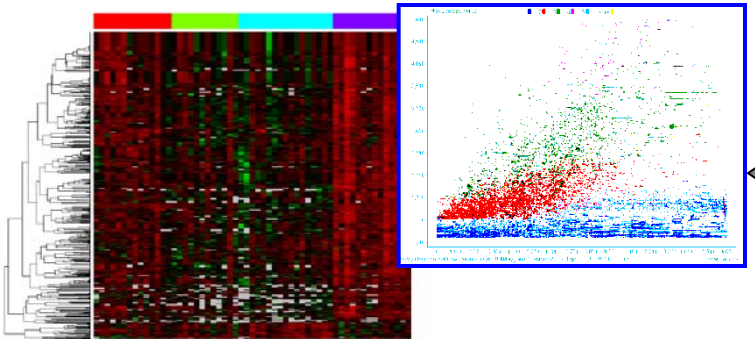
High Throughput Data Analysis: LC-IMS-MS Pipeline



High Throughput Measurement Platforms



Quantitative Proteomics Analysis



Pacific Northwest NATIONAL LABORATORY

Proudly Operated by Battelle *Since 1965*

Interactive Data Analysis: BRM-MeDICi workflows

- User-directed workflows integrate
  - Web queries
  - High performance algorithms
  - User-defined data and input

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Computational Systems Biology at PNNL

- Leadership in HPC applications for
  - High-throughput sequence analysis and feature extraction
  - Graph and network analysis
  - Integrative Modeling and simulation
  - Statistical data analysis
  - High-throughput data analysis
  - Large-scale data management
  - Workflows and visualization tool development
- PNNL is providing leadership in the development of next generation measurement capability

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965