



Dynamic Visualization for Systems Biology and Metaproteomics

Challenge:

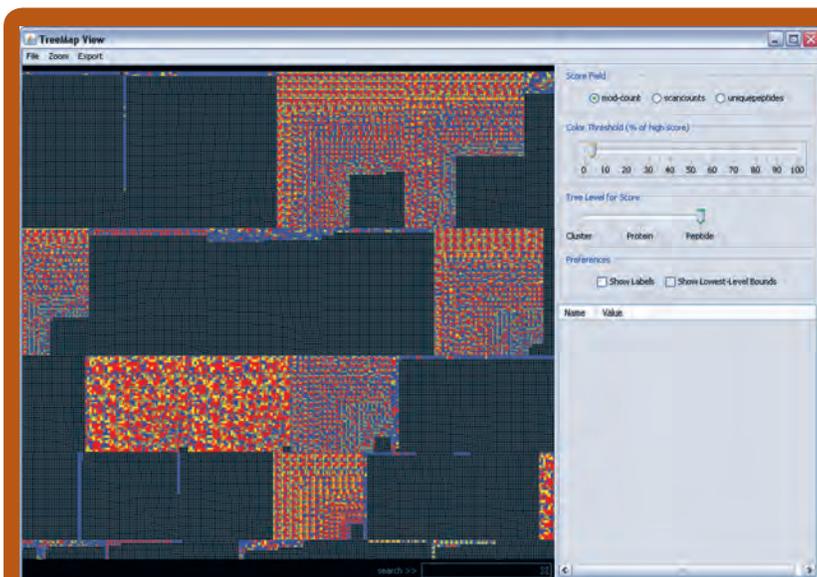
Navigating the massive amounts of data being collected in the field of systems biology and metaproteomics is one of the most pressing technical challenges facing researchers today. This is especially true in the field of comprehensive and quantitative high-throughput microbial metaproteomics, where new tools are needed to effectively analyze data for scientific insight.

Unlike studying isolated sets of proteins, metaproteomics involves the collection and analysis of large volumes of data from protein samples drawn from a specific environmental “community.” By looking at a community, researchers hope to identify changes in proteomic structure and function through the identification of patterns and other visual signs. Researchers believe these visual patterns will serve as predictors for disease and impacts to the environment. However, few practical tools exist for investigators to clearly view, analyze, and explore protein patterns from massive, complex data sets.

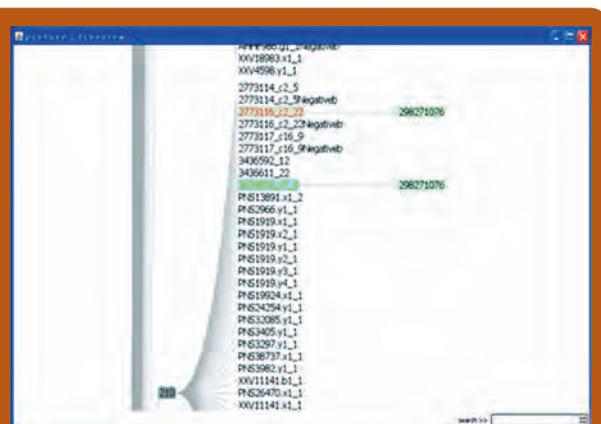
The need for visual analytical tools is especially important to Pacific Northwest National Laboratory (PNNL) Scientist and

Researcher, Angela Norbeck. In collaboration with Oregon State University’s Dr. Stephen Giovannoni and oceanographic researchers from across the nation, Norbeck and the research team are investigating the dynamics of microbial populations in the Sargasso Sea to determine the effects of carbon dioxide on the environment. The study, funded by the National Science Foundation, features one of the largest protein data sets ever compiled. During a four-year period, researchers collected, on a monthly basis, ocean samples comprising millions of proteome data requiring analysis for number, types, and functions within the community. Unfortunately, slow conventional reduction analysis methods can result in loss of information and potentially biased interpretation of results.

To address this challenge, PNNL computational scientists developed a dynamic visualization system, or “toolkit,” that allows researchers to easily navigate and visualize the proteomics data on their desktops, reducing the time of analysis from several weeks to just hours. Here, millions of data points can be categorized in a visual, methodical manner that previously was not possible.



Dynamic TreeMap



Dynamic TreeView

PNNL’s Dynamic Visualization Toolkit Suite allows researchers to visually explore and manipulate protein patterns and groups on the desktop (as depicted in these images), where previous data analysis methods were confined to more conventional text forms such as Excel.



“Advancements being made today in data-intensive computing are very exciting, providing new analytical tools which allow for tables of numbers to suddenly become a visual illustration for more thorough data analysis.”

Angela Norbeck
Scientist

Capability:

A set of visualization tools was provided by PNNL's Dynamic Visualization Toolkit Suite (DVTS), which was partially developed upon the open-source Prefuse visualization toolkit. For the demonstration, the *Dynamic TreeMap* and *Dynamic TreeView* tools from DVTS were specifically applied to visually explore microbial community proteomics data. Integrating visual and text analysis functions enables easy identification of protein groups that change in response to the condition being studied. As dynamic hierarchical visualizations, the *Dynamic TreeMap* and *Dynamic TreeView* allow users to map millions of proteins to sequence clusters, species taxonomies, or hierarchies emerging from user-defined clustering algorithms. The tools have been adapted to accept and incorporate temporal proteomics data and to support specific proteomics analysis paths. The tools also are designed to be modular and self-contained and used as components within automated workflows such as those supported by PNNL's Middleware for Data Intensive Computing (MeDICI) Integration Framework. MeDICI may be used to create and manage data pipelines and to handle large amounts of information with ease—providing scientists with a more scalable, flexible workflow environment that is easier to construct, execute, and navigate.

In analyzing the microbial community proteomics data, the *Dynamic TreeMap* is the core visual analytic tool with hierarchical grids to show patterns within a community of proteins. From the grids, the user can view a summary of the data and choose to hide or filter irrelevant proteins or display those of interest. Within a large sample set, this filtering allows scientists to look for patterns under differing conditions of their choice. To complement the *Dynamic TreeMap*, the *Dynamic TreeView* shows the proteins organized in a horizontal tree structure, providing detailed information, even displaying peptide sequence coverage over proteins, and making the data more explicit and useable. The organization of the *Dynamic TreeView* is similar to the *Dynamic TreeMap*, but with a text-based format that is exportable to other programs.

Impact:

These visual data mining tools allow researchers, such as Norbeck and her Oregon State University partners, to gather significant meaning from the vast amounts of Sargasso Sea data. The tool also is being used by researchers at the University of British Columbia.

The ability to quickly and efficiently visualize and understand large amounts of data is vitally important in the proteomics field as researchers work with increasingly larger samples from multiple sources—water, land, air, and tissue—in the quest to address complex health and environmental issues. With customization, the tools can also be applied to other research domains where huge volumes of data are collected, such as in intelligence, cyber-security, and financial security research.

Learn More:

These visualization tools, and others to complement them, are featured in the article, “Visual Analysis of Dynamic Data Streams,” authored by PNNL's George Chin Jr., Mudita Singhal, Grant Nakamura, Vidhya Gurumoorthi and Natalie Freeman-Cadoret. The article was published in the September 2009 issue of the peer-reviewed journal, *Information Visualization*.

For more information about MeDICI, as well as PNNL's data intensive computing capabilities, please visit the following sites:

- <http://medici.pnl.gov>
- <http://dicomputing.pnl.gov>



To learn more about PNNL tools for dynamic visualization of microbial community proteomics, contact:

George Chin Jr.
Computational Scientist

Pacific Northwest
National Laboratory
P.O. Box 999, MSIN J4-30
Richland, WA 99352
Phone: 509/375-2663
George.Chin@pnl.gov

Pacific Northwest National Laboratory is a Department of Energy Office of Science national laboratory where interdisciplinary teams advance science and technology and deliver solutions to America's most intractable problems in energy, national security, and the environment. PNNL employs 4,650 staff, has a \$954 million annual budget, and has been managed by Ohio-based Battelle since its inception in 1965.



Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by **Battelle** Since 1965