Supercomputing 2008 HPC analytics challenge video entry
Interactive HPC-driver visual analysis for multiple genome datasets

As we progress further into the age of the human genome, biological sequence data is available at an ever-increasing rate. Ten year ago, only a handful of complete genomes were available. Now there are thousands.

To get the most value out of this resource for applications like new medical devices and drugs, renewable energy and environmental cleanup, we must enable researchers to make sense out of this plethora of complex data.

The challenge is to help users identify complex patterns in heterogeneous datasets. High-performance computing gives compute capacity, but we must find ways to integrate this into an iterative process of hypothesis discovery and evaluation.

Our solution is to combine high-performance computing and visual analytics. In this approach, researchers direct high-performance computing resources to calculate relationships between data. The output is mapped to visual metaphors or images that enable scientists to intuitively find patterns and relationships in the data and develop a hypothesis. Once this hypothesis is generated, the compute resources can be used again to evaluate the hypothesis.

For this example, we look at 10 Shewanella species. Shewanella is a microbe that can metabolize some heavy metal pollutants common in DOE waste sites by slowing down their movement in the soil. This is an important possible mechanism for cleaning up legacy waste.

We approach this task by looking at the cellular makeup of the ten different species by focusing on their proteins. Finding all the relationships between proteins in the Shewanella strains involves calculating the similarity between all pairs of approximately 50,000 proteins.

This workflow begins with SHOT, a sensitive homology detection tool. SHOT uses Support Vector Machines to classify uncharacterized proteins against a basis set of well-characterized proteins. This gives an idea of the relative distribution of functions in each species. Distributions are then visualized using Starlight. These visualizations integrate many kinds of data that can be used to sort or filter and refine hypotheses. This visual representation makes it easy to see proteins that are over- or underrepresented in a given species. Proteins of interest are selected and passed on to a second round of high-performance computing.

We use ScalaBLAST to dramatically speedup protein sequence analysis on high-performance computers. The output of this analysis is displayed using visual metaphors or networks that show relationships between the proteins in all ten strains under investigation.

For this demonstration, SHOT was run using 700 processors and completed in 7 minutes. This same task would have taken days on a single workstation. Browsing the SHOT output reveals more than 500 proteins related to Ferric enterobacin receptor, or FepA. This is a protein known to be involved in iron sensing and iron transport in microbes.

The second phase of analysis was performed using 50 processors and completed in less than two minutes. It identified some of these proteins that are also involved in the biogenesis of pili. Pili are appendages required for the formation of biofilms, which are structures composed of many cells held together and protected by an extracellular polymer.

This leads to the hypothesis that proteins involved in pili formation may sense iron in the environment and trigger biofilm formation. This pili biogenesis protein is present in only eight of the ten species. Apparently there has been a gene duplication event resulting in two similar copies in four of the species. Those with two copies may have an enhanced ability to form biofilms in response to iron. The two species that do not appear in the graph may have lost this ability. These hypotheses could be expanded upon via further computational iterations, browsing the SHOT output for additional homology relationships to other proteins related to pili and biofilm biogenesis.

As the cycle continues, computational investigations can lead to follow-on experimental studies and more understanding.

Allowing a scientist to direct high performance computation using this visual analytics pipeline (1) increases the efficiency of the computer resources; (2) it allows real-time interactive hypothesis identification and evaluation and (3) effectively allows users to take advantage of the biological data deluge.

More tightly coupling high-performance computing to the biologist’s workflow maximizes the value of the increasing sequence data for important science efforts aimed at human health, our nation’s security and cleaning the environment.