In the past decade the scientific community has witnessed the most rapid accumulation of gene sequence data and data related to physiology and biochemistry of organisms in the history. To date 237 genomes have been completely sequenced and over a thousand genomes are at different stages of completion. This wealth of genomic information will dramatically accelerate progress toward a comprehensive understanding of the genetic mechanisms involved in diverse biochemical processes pertinent to bioremediation, medicine, biotechnology and agriculture. Crucial to this understanding is the development of tools for the high throughput assimilation, interpretation, and organization of various type of data.

We will demonstrate the functionality and utility of PUMA2 – an integrated Grid technology-based computational environment for high-throughput genetic sequence analysis and evolutionary analysis of metabolism. PUMA2 is being developed at the Mathematics and Computer Science division of Argonne National Laboratory. The current release contains sequence data, annotations and the results of analyses for over 1050 organisms. It enables the interactive analysis of data by a variety of bioinformatics tools, including tools developed by our group (e.g. Chisel, PhyloBlocks) in a framework of metabolic and phenotypic information. The integration of various classes of biological information in PUMA2 allows the identification and analysis of co-evolutionary changes of metabolic pathways, genomic organization, and the study of enzymes characteristic of taxonomic and phenotypic groups of organisms.

The following capabilities of PUMA2 will be discussed during our demonstration:

1.) An environment for high-resolution comparative analysis of genomes. PUMA2 integrates sequence data and annotations from a number of public databases, including RefSeq, PIR, SwissProt, Trembl, metabolic and enzymatic information from EMP database, KEGG, Brenda, Enzyme; phenotypic and taxonomic information. Every sequence in PUMA2 is annotated with information from public resources and the results of pre-computed analysis by a variety of bioinformatics tools including, but not limited to BLAST, InterPro and Blocks. PUMA2 enables interactive analysis of sequence data by conventional bioinformatics tools, including tools developed by our group. These include: Chisel -- a workbench for the identification and characterization of taxonomic and phenotypic versions of enzymes; PhyloBlocks and BloBlast, tools for interactive evolutionary analysis of protein families and the development of HMM models.

2.) Representation of Metabolic pathways in PUMA2. PUMA2 contains metabolic pathways information from the EMP database of Enzymes and Metabolic Pathways. Metabolic pathways are presented in the framework of a general functional overview, providing a hierarchical representation of biological processes. All pathway diagrams are annotated with information regarding chemical compounds from EMP, and provide links to sequence data and annotations. We will introduce the audience to some examples of pathway representations in PUMA2 and demonstrate new tools for comparative analysis of metabolic networks in a framework of phenotypic and taxonomic information being developed by our group. PUMA2 tools for the development of metabolic reconstructions from the sequence data will be also presented.

3.) Automated Grid-technology based computational infrastructure for analysis of genomic data in PUMA2. The grid offers an ideal platform for high-throughput bioinformatics due to its aggregated and distributed computational and storage infrastructure. To utilize the advantages offered by the Grid we have developed GNARE, a scalable computational system for the high-throughput analysis of genomes, which provides the backend for PUMA2. GNARE efficiently automates the major steps of genome analysis, such as the acquisition of data, data analysis, and storage of the results and annotations. The high-throughput computations in GNARE are performed using distributed Grid computing resources. During the demonstration, given adequate resources, we will perform a real time analysis of a bacterial genome by bioinformatics tools (e.g. Blast, Blocks) using GNARE.

Availability: PUMA2 is a public resource, available at http://compbio.mcs.anl.gov/puma2. PUMA2 tools for the interactive evolutionary analysis of protein families PhyloBlocks and BloBlast are assessable from http://compbio.mcs.anl.gov and available for download upon request.