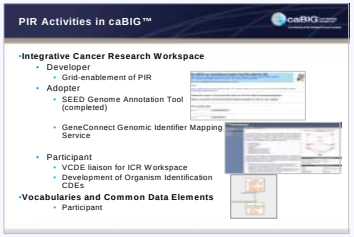
Science Gateway for Protein Analysis on the TeraGrid

Leslie Arminski, Hongzhan Huang, Peter McGarvey, Baris Suzek, Cathy Wu

Protein Information Resource, Lombardi Comprehensive Cancer Center, Georgetown University Medical Center, Washington, DC, 20007

Protein Information Resource (PIR)

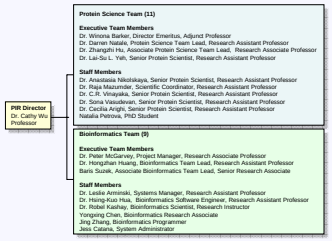


http://cagrid-portal.nci.nih.gov/

http://www.uniprot.org

http://pir.georgetown.edu

http://pir.georgetown.edu/proteomics



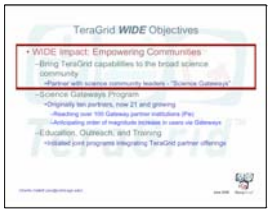
Abstract

The TeraGrid, the largest computational and storage grid in the United States, allows users with common interests to use national resources through science gateways. A Science Gateway enables communities of users associated with a common scientific goal to use TeraGrid resources via a common interface. A current problem in bioinformatics is the large influx of data from genomic and environmental sequencing projects which is beginning to overwhelm the computational resources of many universities and bioinformatics centers. This is particularly true for studies that wish to address global questions sampling all sequence space and not limit the analysis to a particular taxonomic group. The Protein Information Resource (PIR) at Georgetown obtained a development allocation of CPU time on TeraGrid to investigate the practicality of performing large scale bioinformatics calculations on TeraGrid. Here we present lessons learned on our initial tests on running reciprocal BLAST on all protein sequences on the TeraGrid as compared to running locally on our own 50 node (100 cpu) cluster. As a result of this analysis PIR is planning to prototype a science gateway for an automated protein analysis system that will utilize TeraGrid for computationally intensive global analyses. Users would input protein/peptide sequences or a valid protein database ID and receive an integrated knowledge matrix including complete information on gene/peptide to protein database ID mapping, functional analysis including, protein name, family, domain, motif, site, post-translational modification, isoforms, gene ontology, pathways and network discovery. The web interface would allow user interaction to further filter, investigate and analyze the results.

Phase 1 of this project will consume selected caBIG data resources like gridPIR and GeneConnect.
Phase 2 will develop a caBIG silver level compliant analytical service API and caGRID-enable this analytical service.

TeraGrid

- **Largest Computational and Storage Grid in the United States**
- **102 Teraflops Computing Capability**
- **15 Petabytes Online and Archival Storage**
- **Coordinated through Grid Infrastructure Group at the University of Chicago**
- **Resource Partners at Indiana University, Oak Ridge National Laboratory, National Center for Supercomputing Applications, Pittsburgh Supercomputing Center, Purdue University, San Diego Supercomputer Center, Texas Advanced Computing Center, University of Chicago/Argonne National Laboratory, and the National Center for Atmospheric Research**



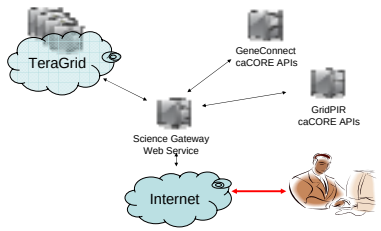
Reciprocal (all-against-all) BLAST Analysis of the UniProt Knowledgebase - Compute Time Comparison

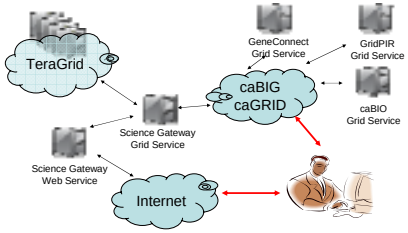
- **Local Resources – (50x2cpu Xeon 2.4GHz) Linux Cluster -Approximately Two Weeks**
- **TeraGrid Resources – (LeMieux at PSC) 6TF (750x4cpu EV68 Alpha) Sierra Cluster - Six Days**

TeraGrid Lessons Learned – It is a Shared Resource

- **Significant Time Savings Are Possible**
- **Queue Times can be long – Especially for Jobs That Consume Large Resources**
- **Caps on the Amount of Resources That Can Be Used During a Defined Period of Time**
- **Caps on the Number of Jobs in the Job Scheduler**
- **Wall Time Limit is Set Too Low for Jobs Running on Many Nodes**
- **Network Glitches May Cause Job Results to be Lost During Transfer from Local Nodes to Storage – Refunds Are Possible**

Science Gateway





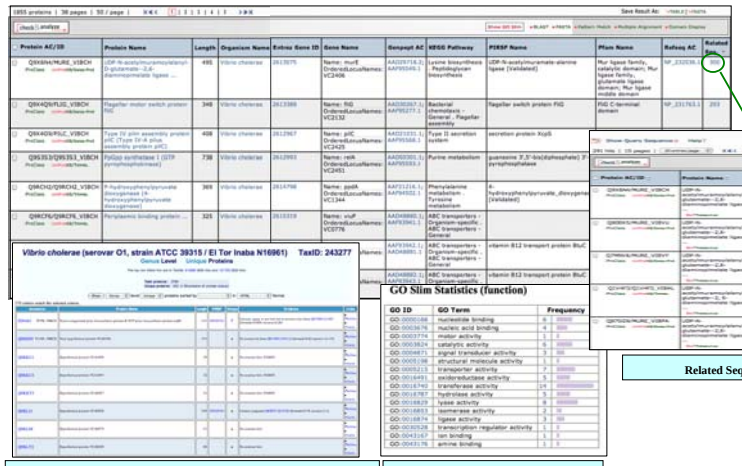
Phase 1: architecture of this project will consume selected caBIG data resources like gridPIR & GeneConnect.

Phase 2: architecture will employ a caBIG silver compliant analytical service API and caGRID-enable this analytical service.

Prototype System

The Science Gateway for Protein Analysis will perform a number of large-scale computations on a regular basis on TeraGrid, provide additional analysis in real time and make the results available to the bioinformatics and medical community. For an initial implementation we propose a modular global protein analysis pipeline consisting of: 1) a reciprocal BLAST analysis to provide pair-wise comparisons of all proteins in UniProtKB; 2) a global analysis to find all known functional domains, sites and families using InterProScan; and, 3) a global scan for all signal peptides and transmembrane domains, using Phobius.

Prototype web interface of the integrated knowledge matrix, including complete information on gene/peptide to protein database ID mapping, functional analysis including, protein name, family, domain, motif, site, post-translational modification, isoforms, gene ontology, pathways and network discovery. The web interface would allow users to further filter, investigate and analyze the results.



Genome Comparisons, derived from Reciprocal BLAST Analysis

GO Analysis

Related Sequences, derived from Reciprocal BLAST Analysis