ESPP Computational Core Overview

Paramvir S. Dehal[1,2], Morgan N. Price[1,2], Eric J. Alm[1,3], Dylan Chivian[1,2], Katherine H. Huang[1,2], Y. Wayne Huang[1,2], Marcin P. Joachimiak[1,2], Keith Keller[1,2], Adam P. Arkin[1,2,4,5]

[1]Virtual Institute for Microbial Stress and Survival, http://vimss.lbl.gov; [2]Lawrence Berkeley National Laboratory;[3]Department of Biological Engineering, MIT; [4]University of California, Berkeley and [5]Department of Bioengineering, University of California, Berkeley

ESPP2 Environmental Stress Pathway Project

DOE GENOMICS:GTL ACCELERATING DISCOVERY FOR ENERGY AND ENVIRONMENT — OFFICE OF SCIENCE U.S. DEPARTMENT OF ENERGY

VIMSS Virtual Institute for Microbial Stress and Survival — Massachusetts Institute of Technology — University of Missouri — University of Washington — Montana State University — Southern Illinois University Carbondale — University of Oklahoma — Sandia National Laboratories — Oak Ridge National Laboratory, Managed by UT Battelle for the Department of Energy — Berkeley, University of California — 75 Years of World Class Science 1931-2006

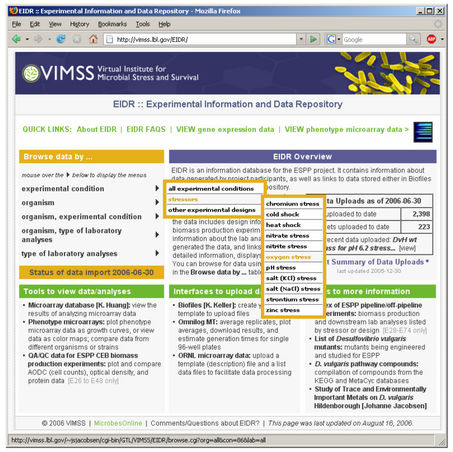
http://vimss.lbl.gov/

INTRODUCTION

The VIMSS Computational Core group is responsible for data management, data integration, data analysis, and comparative and evolutionary genomic analysis of the data for the VIMSS project. We have expanded and extended our existing tools sets for comparative and evolutionary genomics and microarray analysis as well as creating new tools for our proteomic and metabolomic data sets. Our analysis has been incorporated into our comparative genomics website MicrobesOnline (http://www.microbesonline.org) and made available to the wider research community. By taking advantage of the diverse functional and comparative datasets, we have been able to pursue large evolutionary studies.

Data Management

All data generated by ESPP continues to be stored in our Experimental Information and Data Repository (http://vimss.lbl.gov/EIDR/). Researches have access to datasets from biomass production, growth curves, image data, mass spec data, phenotype microarray data and transcriptomic, proteomic and metabolomic data. New functionality has been added for storage of information relating to mutants and protein complex data, in addition to new visualization for assessing existing data sets such as the phenotype microarrays.



MicrobesOnline

The MicrobesOnline database (http://www.microbesonline.org) currently holds 705 microbial genomes and is updated quarterly with the NCBI RefSeq database, providing an important comparative genomics resource to the community. New functionality added this year includes the addition of a phylogenetic tree based genome browser that allows users to view their genes and genomes of interest within an evolutionary framework, tools allowing users to search for novel regulatory binding site motifs or matches to existing regulatory binding motifs across a user selected set of genes using our Gene Carts, tools to compare multiple microarray expression data across genes and genomes and integration with the RegTransDb of experimentally verified regulatory binding sites.

MicrobesOnline continues to provide an interface for genome annotation, which like all the tools reported here, is freely available to the scientific community. To keep up with the rapidly expanding set of sequenced genomes, we have begun to investigate methods for accelerating our annotation pipeline. In particular we have completed methods to speed up the most time consuming process, homology searching through HMM alignments and all against all BLAST. We are now in the process of accelerating phylogenetic tree building. Over the next year we will be releasing methods that will allow us to deal with the many millions of gene sequences generated from metagenomics.

Over the next year, several new features will be added to the MicrobesOnline resource. In addition to expansion of our Microarray Expression datasets and analysis tools, we will begin to present Metabolite and Metabolomics data through our Pathway tools.

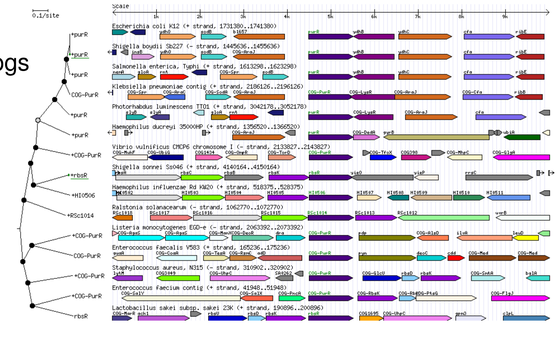
MicrobesOnline TreeBrowser

Sequence similarity search is the first step in computational gene characterization. However, this process currently generates approximately 1,000 significant matches per gene. Sorting through this large data set can be tedious and time consuming as many of the top matches are to genes from closely related strains and most of the alignments are to gene families. Without knowing the species relationships, the similarity matches alone will not reveal the genes distribution or evolutionary history.

Our solution is the MicrobesOnline TreeBrowser. A phylogenetic tree is constructed for each gene family on the basis of domains, COGs or ad-hoc BLAST families. The TreeBrowser can then display your gene within this framework, highlighting the characterized homologs, collapsing branches of close relatives allowing more distant relationships to be shown and reconciling the gene tree with the species tree to elucidate the evolutionary history.

E. coli purR:

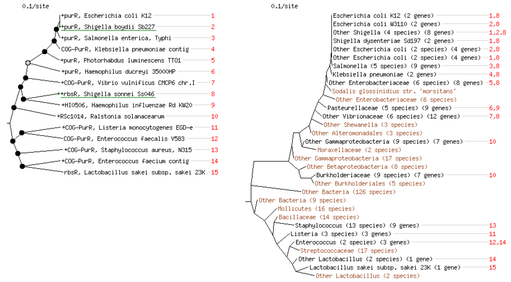
the 93 closest homologs



Gene Tree

versus

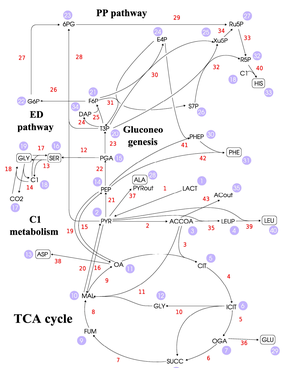
Species Tree



Metabolomics & Metabolic Flux

The ESPP Functional Genomics Core has begun developing methods for the high throughput identification of metabolites from cell extract and for determining the metabolic flux through central metabolism using ^{13}C isotopomer analysis.

To support this effort, we are focusing on creating the computational infrastructure necessary to automate the data acquisition process by working with the experimentalists to remove manual data processing. Additionally, we are beginning to create the tools for analysis and interpretation of the data within the MicrobesOnline Pathway toolset.



Fast Sequence Search

MicrobesOnline currently has 705 microbial genomes with nearly 2 million genes, soon we will have several thousands of genomes and gigabases of sequence from metagenomics and isolate sequencing. Sequence similarity determination using BLAST and domain identification using HMMer are quickly become possible only on large compute clusters.

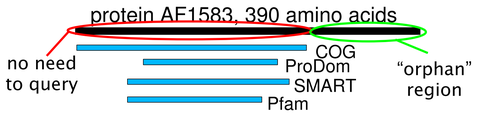
HMMFast

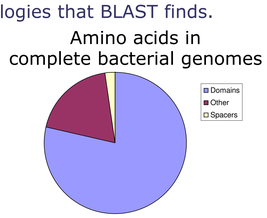
We achieve a 50-100x speed up, with >98% sensitivity, over the traditional HMMer search by combining a sensitive PSI-BLAST search with HMMer to remove false positives.

Database	# HMMs	# Test Genomes	HMMer CPU hrs	PSI-BLAST CPU hrs	PSI-BLAST % with hits	Fast hrs (predicted)	% of HMMer Hits Missed
Pfam	8,286	10	174	0.17	0.25%	0.60	1.9
TIGRfam	2,946	10	116	0.33	0.70%	1.14	1.4
Supfam	10,894	5	83	0.04	0.38%	0.36	1.7

FastBLAST

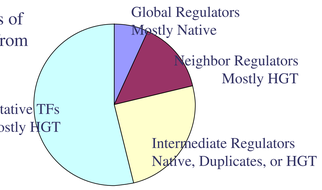
FastBLAST masks known domains and uses CD-HIT to further reduce both the query size and the database size. FastBLAST is ~18x faster than the traditional All-Against-All BLAST, and finds 99.8% of the homologies that BLAST finds.





Transcription Factor Evolution

Evolutionary histories of transcription factors from *Escherichia coli* K12



- Duplications of TFs within the *E. coli* lineage are rare
  - most paralogs arose by HGT
- Regulatory network rarely expands by gene duplication (7% of network)
  - regulatory similarities between paralogs are often convergent (17/30)
- Regulatory sites mostly arise after genes are acquired by HGT
  - but in neighbor regulation, operon and TF often acquired together
  - sites for other regulators (e.g. CRP) sometimes transferred too
- HGT genes are under more complex regulation
  - regulation evolves rapidly under strong selection
  - CRP preferentially regulates HGT genes, & most CRP sites evolved recently
- Co-HGT of neighbor regulators allows prediction
  - Most uncharacterized TFs are neighbor regulators?

ACKNOWLEDGEMENT

ESPP2 (MDCASE) is part of the Virtual Institute for Microbial Stress and Survival (VIMSS) supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics:GTL Program through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U. S. Department of Energy.