VIMSS Computational Core

Paramvir S. Dehal[1,2*] (PSDehal@lbl.gov), Eric J. Alm[1,3], Dylan Chivian[1,2,4], Katherine H. Huang[1,2], Marcin P. Joachimiak[1,2], Keith Keller[1,2], Morgan N. Price[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, Berkeley, CA, 94720; [3]Department of Biological Engineering, MIT, Cambridge, MA, 02139; [4]DOE Joint BioEnergy Institute, Emeryville, CA; and [5]Department of Bioengineering, University of California, Berkeley, CA, 94720

Acknowledgements
This work was part of the Virtual Institute for Microbial Stress and Survival (http://VIMSS.lbl.gov) 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

Background: 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 functional genomics to deal with new data produced by the VIMSS ESPP2 members. The Computational Core is developing methods to store and analyze diverse data sets including: microarrays, ChIP-chip arrays, tiling arrays, proteomics, metabolomics, metabolic flux, phylochips, metagenomics sequencing, genome sequencing, growth curves, phenotype arrays, knock out strain collections and links to existing literature and web based resources. Our analysis has been incorporated into our comparative and functional genomics website MicrobesOnline (http://www.microbesonline.org) and made available to the wider research community. By taking advantage of data integration across diverse functional and comparative datasets, we have been able to pursue large research projects in evolutionary and systems biology studies.

Data Integration: Data management, integration and distribution are critical functions for all large projects. A primary goal of the Computational Core is to capture all experimental data from the ESPP2 investigators, including relevant metadata, raw data and processed data, and to make these data sets available through intuitive queries. Our group has developed Experimental Information and Data Repository (http://vimss.lbl.gov/EIDR/) and the MicrobesOnline database to provide this functionality. 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 mutant strains, transposon knockout libraries and protein complex data, in addition to new visualization for assessing existing data sets such as the phenotype microarrays.

Data Analysis: The Computational Core has focused on using the data generated by the ESPP2 project to understand the stress response of *Desulfovibrio vulgaris* Hildenbourgh. We work closely with the other core groups within the ESPP2 project to assist in data analysis. Over the past year, this has included co-culture laboratory evolution, 16S barcode data, and phenotype analysis for the Applied Environmental Microbiology Core (AEMC) and transcriptomic, tiling

arrays, and metabolite analysis for the Functional Genomics Core (FGC). New research being pursued this year by the Computational Core includes: development of new methods for data compendium analysis using biclustering which combines transcriptomic, proteomic, interaction and gene neighborhood data in order to predicted regulatory structures; computational predictions of amino acid synthesis pathways in DvH (working closely with the FGC to verify predictions); evolutionary analysis of lineage specific gene expansion across bacteria; sub N squared phylogenetic tree reconstruction by developing the FastTree program; and large scale comparative metagenomic sequence analysis.

The MicrobesOnline Database: The MicrobesOnline database (http://www.microbesonline.org) currently holds over 1000 microbial genomes and will be updated semi-annually, providing an important comparative and functional genomics resource to the community. New functionality added this year includes the addition of fungal genomes and the framework for adding additional eukaryotic genomes, an updated user interface for the phylogenetic tree based genome browser that allows users to view their genes and genomes of interest within an evolutionary framework, improved tools to compare multiple microarray expression data across genes and genomes, phylogenetic profile searches using our high quality species tree, and addition of external microarray data from the Many Microbial Microarrays Database for bacteria and Yeast. Additionally we have begun adding metagenomic data to MicrobesOnline.

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 work FastHMM and FastBLAST, methods to speed up the most time consuming process of our analysis pipeline, homology searching through HMM alignments and all against all BLAST. These methods now enable us to deal with the many millions of gene sequences generated from metagenomics. And our FastTree program allows us to create phylogenetic trees for all gene families, even those with over 100,000 members, so that all genes can be studied within an evolutionary framework.