MicrobesOnline: an integrated portal for comparative functional genomics

Marcin P. Joachimiak[1,2], Katherine H. Huang[1,2], Eric J. Alm[1,3], Dylan Chivian[1,2], Paramvir S. Dehal[1,2], Y. Wayne Huang[1,2], Janet Jacobsen[1,4], Keith Keller[1,4], Morgan N. Price[1,2], Adam P. Arkin[1,2,4,5,6]

[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]University of California, Berkeley, CA, 94720; [5]Howard Hughes Medical Institute and [6]Department of Bioengineering, University of California, Berkeley, CA, 94720

ESPP Environmental Stress Pathway Project

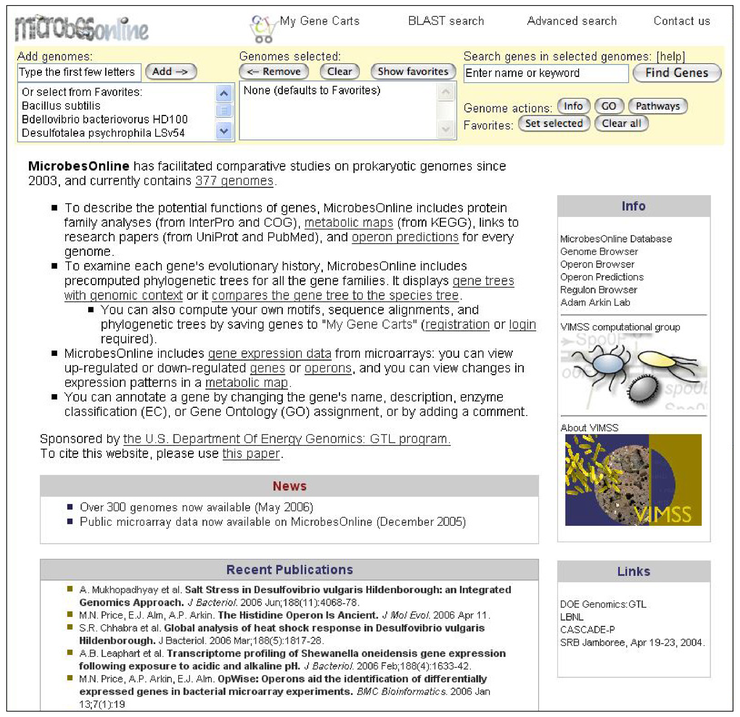
DOE GENOMICS:GTL ACCELERATING DISCOVERY FOR ENERGY AND ENVIRONMENT — OFFICE OF SCIENCE U.S. DEPARTMENT OF ENERGY

VIMSS Virtual Institute for Microbial Stress and Survival · Missouri · University of Washington · Montana State University · Oak Ridge National Laboratory · Sandia National Laboratories · Berkeley University of California · 75 Years of World Class Science

http://vimss.lbl.gov/

Introduction

The Virtual Institute for Microbial Stress and Survival (VIMSS, http://vimss.lbl.gov) funded by the Dept. of Energy's Genomics:GTL Program, is dedicated to using integrated environmental, functional genomic, and comparative sequence and phylogeny data to understand mechanisms by which microbes survive in uncertain environments while carrying out processes of interest for bioremediation and energy generation. To support this work, VIMSS has developed a Web portal with an underlying database and analyses for comparative functional genomics of bacteria and archaea. Since 2003 MicrobesOnline (http://www.microbesonline.org) has been enabling comparative genome analysis and currently includes 465 complete genomes, of which 423 are microbial, and offers a suite of analysis and tools including: a multi-species genome browser, operon and regulon prediction methods and results, a combined gene and species phylogeny browser, a gene ontology browser, a workbench for sequence analysis (including sequence motif detection, motif searches, sequence alignment and phylogeny reconstruction), and capabilities for community annotation of genomes.



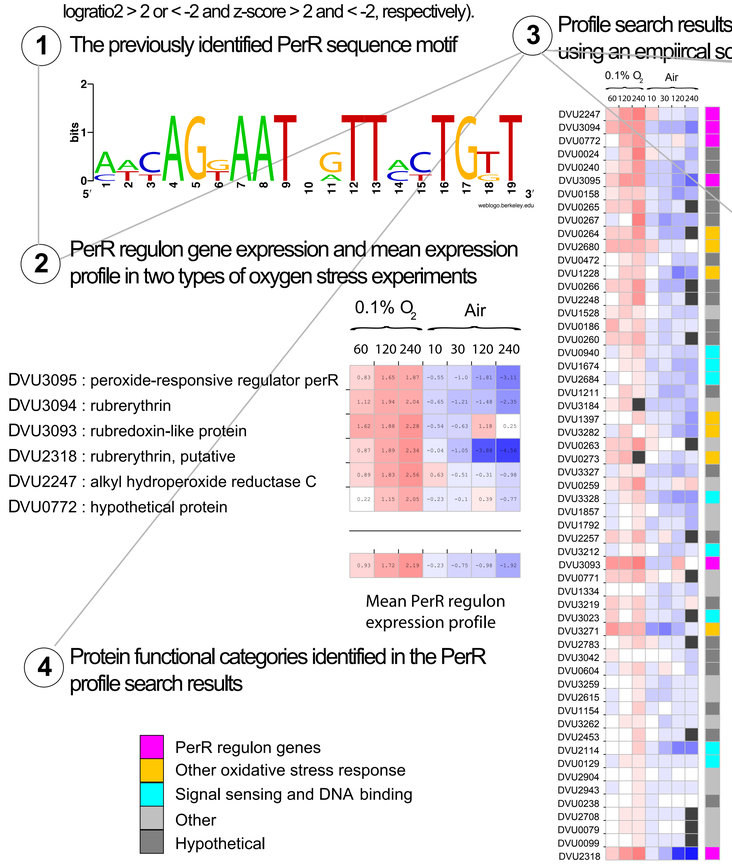
VIMSS integrates functional genomic data and provides novel web-based viewing and mining tools for gene expression microarray, proteomic, and phenotype microarray data. Currently, these data are mostly project generated for wild-type and mutants of *Desulfovibrio vulgaris* and *Shewanella oneidensis* exposed to stress conditions found at DOE field sites. Selecting an organism or gene of interest in MicrobesOnline leads to information about and data viewers for VIMSS experiments conducted on that organism and involving that gene or gene product. It is also now possible to view microarray data from multiple stress conditions as an interactive heatmap and to analyze correlations between gene expression results from different experiments. Among the major new features is the ability to search any subset of experiments in the microarray data compendium for similar gene matches to a mean expression profile derived from an a priori determined group of genes (e.g., a known or predicted regulon). These new compendium-wide functionalities allow one to observe patterns in gene expression changes across multiple conditions and to search for similarities to these patterns.

The MicrobesOnline microarray data compendium

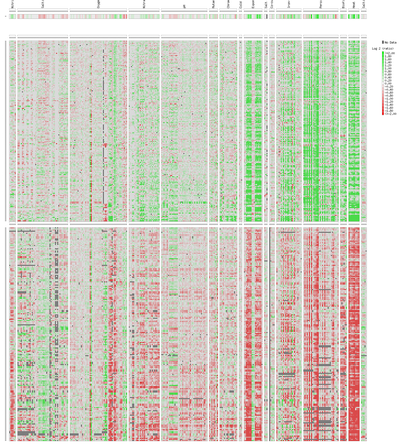
Stress	Experiment groups	Experiment comparisons	Hybridizations
Oxygen	7	47	128
Co-culture	7	7	98
Salt:NaCl	7	52	134
Nitrate	5	28	93
pH	4	42	112
Iron-impact	4	55	266
Iron-limited	3	23	54
Biofilm	3	6	30
Peroxide	2	32	72
Exponential VS Stationary	2	17	50
Chromium	2	16	54
Mutant	2	8	24
Cold shock	1	3	17
Heat shock	1	10	33
Salt Adaptation	1	3	16
Sulfur-limited	1	N/A	2
Salt:KCl	1	5	30
Total	**53**	**354**	**1213**

Gene-gene expression profile searches

Rodionov et al (Genome Biology 2003) identified a putative PerR (peroxide response) regulon in the Desulfovibrio vulgaris Hildenborough genome using comparative genomics and sequence motif detection. Subsequently, a mild environmental stress microarray experiment, microaerobic stress (0.1 % O2), identified members of this regulon amongst the only 12 significantly changing genes (total of genes with logratio2 > 2 or < -2 and z-score > 2 and < -2, respectively).

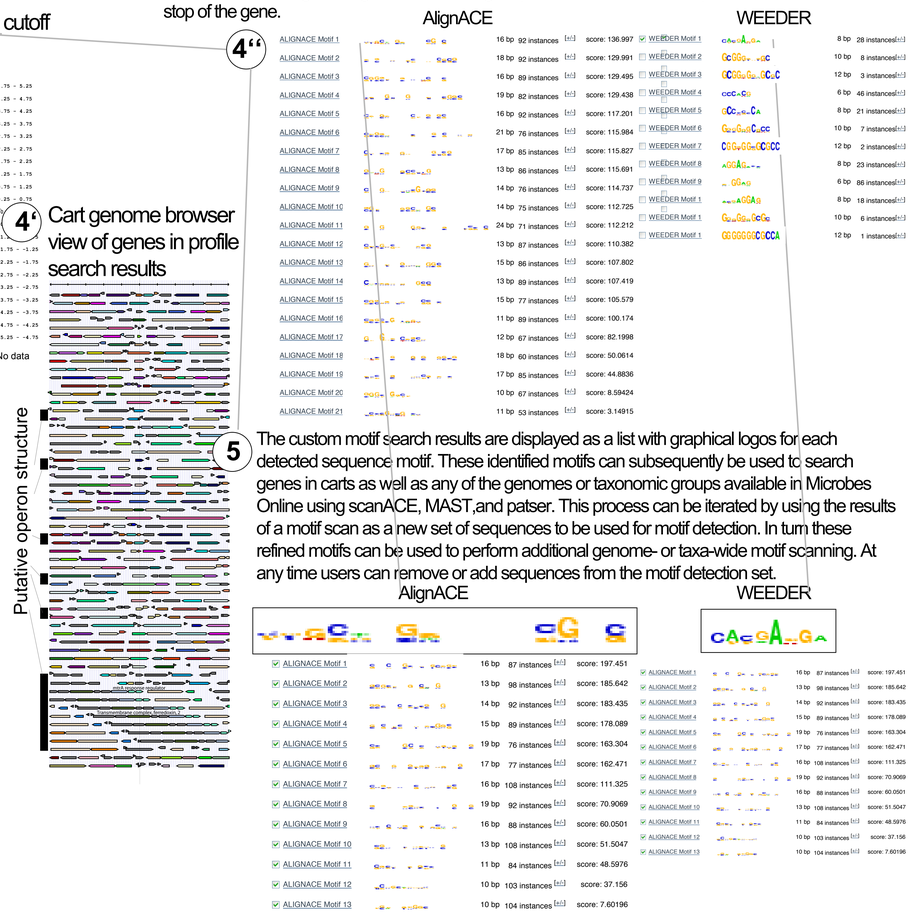


Compendium gene-gene expression profile searches



Custom and iterative regulatory sequence motif detection and scanning

MicrobesOnline includes suite of sequence motif detection tools. Starting from user specified genes in a user's cart, the motif detection interface allows creating custom motif search jobs with the AlignACE, MEME, or WEEDER programs. The interface allows specifying the regions to be searched, i.e., upstream and downstream regions of genes, or in coding sequence, with coordinates specified relative to the start or stop of the gene.



Conclusions

Gene expression data and regulatory sequence motif detection in of themselves can be powerful tools for generating molecular function and system hypothesis. Moreover, these methods can be combined to provide additional support for a sequence motif or expression profile similarity alone. Work is in progress on incorporating additional publicly available microarray experiment datasets from a wider taxonomic sampling of microorganisms, enabling comparative phylogeny methods in the context of sequence motifs and expression profiles. In the face of a growing number of datasets and computational methods, MicrobesOnlines serves an important integrative role at both the experimental data and the computational biology software levels. We invite new and old users to try out our new web enabled computational tools and rich microbial dataset environment.

References

*Alm EJ, Huang KH, Price MN, Koche RP, Keller K, Dubchak IL, Arkin AP.
The MicrobesOnline Web site for comparative genomics.
Genome Res. 2005 Jul;15(7):1015-22.

*Bailey TL, Elkan C
Fitting a mixture model by expectation maximization to discover motifs in biopolymers.
PSIC on ISMB, pp. 28-36, AAAI Press, Menlo Park, California, 1994.

*Bailey TL, Gribskov M.
Combining evidence using p-values: application to sequence homology searches.
Bioinformatics, Vol. 14, pp. 48-54, 1998.

*Hertz GZ, Stormo GD.
Identifying DNA and protein patterns with statistically significant alignments of multiple sequences.
Bioinformatics. 1999 Jul-Aug;15(7-8):563-77.

*Mukhopadhyay A, Joachimiak MPP, Redding AMT, Arkin AP, Borglin S, Dehal PS, Chakraborty R, Geller JT, Giles B, Hazen TC, He Q, Joyner DC, Martin VJJ, Wall JD, Yang ZK, Zhou J, Keasling JD.
Comparison of Desulfovibrio vulgaris Hildenborough response to microaerobic and aerobic exposure.
Submitted.

*Pavesi G, Mereghetti P, Mauri G, Pesole G.
Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes.
Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W199-203.

*Pavesi G, Mauri G, Pesole G.
An algorithm for finding signals of unknown length in DNA sequences.
Bioinformatics. 2001;17 Suppl 1:S207-14.

*Rodionov DA, Dubchak I, Arkin A, Alm E, Gelfand MS.
Reconstruction of regulatory and metabolic pathways in metal-reducing delta-proteobacteria.
Genome Biol. 2004;5(11):R90. Epub 2004 Oct 22.

*Roth FP, Hughes JD, Estep PW, Church GM.
Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation.
Nat Biotechnol. 1998 Oct;16(10):939-45.

Acknowledgment

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