Final Scientific/Technical report

DOE Award number:	DE-FG02-07ER64388
Name of recipient:	Current PI: Dr. Daniel Segrè (since April 2009). Initially Awarded to: Dr. Timothy Gardner.
Project title:	Integrated genome-based studies of Shewanella ecophysiology
Project director/PI:	Prof. Jim Fredrickson (PNNL, Richmond, WA)
Consortium/Team members:	Shewanella federation
Patentable material/Protected data:	N/A

Executive summary:

This project was a component of the *Shewanella* Federation and, as such, contributed to the overall goal of applying the genomic tools to better understand eco-physiology and speciation of respiratory-versatile members of *Shewanella* genus. Our role at Boston University was to perform bioreactor and high throughput gene expression microarrays, and combine dynamic flux balance modeling with experimentally obtained transcriptional and gene expression datasets from different growth conditions. In the first part of project, we designed the *S. oneidensis* microarray probes for Affymetrix Inc. (based in California), then we identified the pathways of carbon utilization in the metal-reducing marine bacterium *Shewanella oneidensis* MR-1, using our newly designed high-density oligonucleotide Affymetrix microarray on *Shewanella* cells grown with various carbon sources. Next, using a combination of experimental and computational approaches, we built algorithm and methods to integrate the transcriptional and metabolic regulatory networks of *S. oneidensis*. Specifically, we combined mRNA microarray and metabolite measurements with statistical inference and dynamic flux balance analysis (dFBA) to study the transcriptional response of *S. oneidensis* MR-1 as it passes through exponential, stationary, and transition phases. By measuring time-dependent mRNA expression levels during batch growth of *S. oneidensis* MR-1 under two radically different nutrient compositions (minimal lactate and nutritionally rich LB medium), we obtain detailed snapshots of the regulatory strategies used by this bacterium to cope with gradually changing nutrient availability. In addition to traditional clustering, which provides a first indication of major regulatory trends and transcription factors activities, we developed and implemented a new computational approach for Dynamic Detection of Transcriptional Triggers (D2T2). This new method allows us to infer a putative topology of transcriptional dependencies, with special emphasis on the nodes at which external stimuli are expected to affect the internal dynamics. In parallel, we addressed the question of how to compare transcriptional profiles across different time-course experiments. Our growth derivative mapping (GDM) method makes it possible to relate

with each other points that correspond to the same relative growth rate in different media sets. This mapping allowed us to discriminate between genes that display an environment-independent behavior, and genes whose transcription seems to be tuned by specific environmental factors. Our analysis highlighted the importance of some specific pathways, whose metabolic relevance was confirmed by dynamic flux balance analysis (dFBA) calculations. In particular, we found that oxygen limitation potentially triggers the activation of genes previously shown to be relevant for anaerobic respiration, and that nitrogen limitation is coupled to storage of glycogen. Both observations have been corroborated by measurement of relevant intracellular and extracellular metabolites, as well as by complementary analyses of literature information and competitive fitness assay data. The pipeline of experimental and computational approaches applied and developed for this work could be extended to other microbes and additional conditions.

Summary of project activities:

Driscoll et al, 2007: To identify pathways of carbon utilization in the metal-reducing marine bacterium *Shewanella oneidensis* MR-1, we assayed the expression of cells grown with various carbon sources using a high-density oligonucleotide Affymetrix microarray. Our expression profiles reveal genes and regulatory mechanisms which govern the sensing, import, and utilization of the nucleoside inosine, the chitin monomer N-acetylglucosamine, and a casein-derived mixture of amino acids. Our analysis suggests a prominent role for the pentose-phosphate and Entner-Doudoroff pathways in energy metabolism, and regulatory coupling between carbon catabolism and electron acceptor pathways. In sum, these results indicate that *S. oneidensis* possesses a broader capacity for carbon utilization than previously reported, a view with implications for optimizing its role in microbial fuel cell and bioremediative applications.

Faith et al. 2008: Many Microbe Microarrays Database (M3D) is designed to facilitate the analysis and visualization of expression data in compendia compiled from multiple laboratories. M3D contains over a thousand Affymetrix microarrays for *Escherichia coli*, *Saccharomyces cerevisiae* and *Shewanella oneidensis*. The expression data is uniformly normalized to make the data generated by different laboratories and researchers more comparable. To facilitate computational analyses, M3D provides raw data (CEL file) and normalized data downloads of each compendium. In addition, web-based construction, visualization and download of custom datasets are provided to facilitate efficient interrogation of the compendium for more focused analyses. The experimental condition metadata in M3D is human curated with each chemical and growth attribute stored as a structured and computable set of experimental features with consistent naming conventions and units. All versions of the normalized compendia constructed for each species are maintained and accessible in perpetuity to facilitate the future interpretation and comparison of results published on M3D data. M3D is accessible at http://m3d.bu.edu/.

Beg et al submitted: we performed a system-level analysis of *S. oneidensis* batch culture, combining time-course gene expression, computational analyses and simulations, and metabolite measurements. We started by performing microarray measurements along the transition from exponential to stationary phase in minimal and rich media. By comparing time-rescaled transcriptional profiles across these two conditions we identified processes

whose regulation is similar in minimal and rich media. Conversely, to find nodes in the genetic network that are heavily influenced by environmental factors, we implemented a new algorithm (Dynamic Detection of Transcriptional Triggers, or D2T2), aimed at finding transcriptional changes that cannot be explained solely by internal network dynamics. D2T2 identifies known and novel regulators that respond to carbon, nitrogen and oxygen limitation. These transcriptional changes suggest an interesting sequence of physiological responses, including a potential coupling between nitrogen depletion and glycogen metabolism, partially recapitulated through dynamics flux balance analysis, and experimentally confirmed by metabolite measurements.

Reznik et al., in preparation: We developed a new algorithm named TEAM (Temporal Expression-based Analysis of Metabolism) that integrates for the first time time-dependent expression data with dynamic Flux Balance Analysis (dFBA). The approach is an extension of a previous algorithm for using gene expression in conjunction with Flux Balance Analysis (GIMME, Gene Inactivity Moderated by Metabolism and Expression). TEAM simulates growth of an organism over time by trying to match gene expression patterns that vary for each gene over a series of multiple time points. The result is a prediction of growth and metabolic behavior that is directly informed by measured gene expression, which infuses the otherwise generic, stoichiometric-based simulation with a specific environmental context. We applied TEAM to our own data compendium associated with batch growth of *Shewanella oneidensis* MR-1. (see project by Beg at al. above). Using TEAM, we were able to rationalize computationally some otherwise unexplainable observations about secreted metabolite. In addition, TEAM allowed us to assess critically the advantages and limitations of integration of flux balance models with expression data.

Journal Publications and Conference presentations:

Journal Publications

Driscoll ME, Romine MF, Juhn FS, Serres MH, McCue LA, Beliaev AS, Fredrickson JK, Gardner TS (2007) Identification of diverse carbon utilization pathways in Shewanella oneidensis MR-1 via expression profiling. Genome Informatics 18: 287-298

Faith JJ, Driscoll ME, Fusaro VA, Cosgrove EJ, Hayete B, Juhn FS, Schneider SJ, Gardner TS (2008) Many Microbe Microarrays Database: Uniformly normalized affymetrix compendia with structured experimental metadata. Nucleic Acids Research 36: D866-D870

Fredrickson JK, Romine MF, Beliaev AS, Auchtung JM, Driscoll ME, Gardner TS, Nealson KH, Osterman AL, Pinchuk G, Reed JL, Rodionov DA, Rodrigues JL, Saffarini DA, Serres MH, Spormann AM, Zhulin IB, Tiedje JM (2008) Towards environmental systems biology of Shewanella. Nat Rev Microbiol 6: 592-603

Klitgord N. and Segre' D., Environments that induce synthetic microbial ecosystems, PLoS Computational Biology (2010), 6(11): e1001002.

Beg QK, Zampieri M, Baldwin S, Klitgord N, Altafini C., Serres MH, Segre D. : Detection of transcriptional triggers in the dynamics of microbial growth: application to a respiratory-versatile bacterium. Revised version in preparation for Nucleic Acids Research.

Reznik E, Baldwin S, Segrè D., Integrating time-series gene expression measurements with a flux model to examine metabolic potential: a case study using Shewanella Oneidensis, in preparation for PLoS Computational Biology.

Conference Presentations

Reznik E, Baldwin S, Beg QK, Segre D. (2011). Deciphering the growth dynamics of Shewanella oneidensis by integrating metabolite and gene expression profiles with stoichiometric modeling. The 2011 Genomic Science Contractor-Grantee meeting/ UDSA-DOE Plant feedstock Genomics for Bioenegy Awardees Meeting, Arlington, VA, Apr. 10-13, 2011. (poster presentation)

Beg QK, Zampieri M, Baldwin S, Klitgord N, Serres MH, Altafini C, Segre D. (2010). Transcriptional dynamics of Shewanella oneidensis MR-1 during growth on minimal and rich media. The 16th Annual Boston Bacterial Meeting (BBM 2010), Harvard University Science Center, Cambridge, MA, June 17-18, 2010.

Beg QK, Zampieri M, Baldwin S, Klitgord N, Serres MH, Altafini C, Segre D. (2010). Identifying the mediators of environmental changes through integration of steady state and time-course gene expression profiles in Shewanella oneidensis MR-1. The 2010 Genomic Science Contractor-Grantee and Knowledge Workshop, sponsored by US-DOE Office of Biological Environment Research, Arlington, VA, Feb. 7-10, 2010. (both poster and oral presentations)

Beg QK, Zampieri M, Klitgord N, Baldwin S, Gardner TS, Segre D. (2009). Systems approach for identifying transcriptional and metabolic changes in Shewanella oneidensis MR-1 during growth-phase transitions. The 10th International Conference on Systems Biology (ICSB-2009), Aug. 30-Sep. 4, 2009, Stanford, CA. (poster presentation)

Zampieri M*, Klitgord N, Beg QK, Altafini C, Segre D. (2009). Inferring drug targets from Gene-Networks. 2nd Conference on Drug Development for the Third World. From Computational Molecular Biology to Experimental Approaches. June 1-5, 2009, Trieste, Italy (*presenting author)

Beg QK, Klitgord N, Zampieri M, Gardner TS and Segre D. (2009). Experimental and computational analysis of growth-phase dependent transcriptional programs in Shewanella oneidensis. Genomics: GTL Contractor-Grantee Workshop VII and USDA-DOE Plant Feedstock Genomics for Bioenergy Awardee Workshop, Bethesda, MD. Feb. 8-11, 2009. (poster presentation)

Beg QK, Klitgord N, Byrne DK, Gardner TS and Segre D. (2008). Growth-phase dependent transcriptional regulation in Shewanella oneidensis MR-1. Shewanella Federation Meeting Fall 2008, Asilomar Conference Grounds, Pacific Grove, CA, Oct. 5-7, 2008. (oral presentation)

Beg QK, Driscoll ME, Cosgrove EJ, Byrne DK, Gardner TS and Segre D. (2008). Regulation of transcriptional programs in Shewanella oneidensis MR-1 in wide range of environmental conditions. Society for Industrial Microbiology's (SIM) Annual Meeting and Exhibition, San Diego, CA, Aug. 10-14, 2008. (poster presentation)

Beg QK, Byrne DK, Driscoll ME, Juhn FS, Shen Y, Faith JJ, Paschalidis I, Segre D and Gardner TS. (2008). Integrated understanding of the metabolic and gene regulatory systems of Shewanella oneidensis MR-1. 108th Annual Meeting of the American Society of Microbiology (ASM), Boston, MA, June 1-5, 2008. (poster presentation)

Byrne DK*, Beg QK, Driscoll ME, Juhn FS, Shen Y, Faith JJ, Paschalidis I, Segre D and Gardner TS. (2008). Systematic identification of regulatory mapping and optimal metabolic engineering strategies in Shewanella oneidensis MR-1. Joint Genomics: GTL Awardee Workshop VI, Metabolic Engineering for Bioenergy Awardee Workshop and Shewanella Federation Meeting, Bethesda, MD. Feb. 9-13, 2008. (*presenting author)

Website reflecting results/data of project/Databases:

1. http://m3d.bu.edu (Many microbes array database holding gene expression data)
2. http://www.ncbi.nlm.nih.gov/geo/ (Accession number GPL8434 for Shewanella oneidensis platform, BU_Shewanella_100k_v1.0)
3. http://shewybase.bu.edu/

Networks or collaborations fostered:

1. Dr. Mattia Zampieri and Dr. Claudio Altafini: SISSA (International School of Advanced Studies), 34136 Trieste, Italy [Participated in the development of new method for identifying transcriptional triggers from gene expression data]
2. Dr. Margrethe H. Serres: Marine Biological Laboratory, Woods Hole, MA, 02543 USA [Participated in biological analysis interpretation of Shewanella expression data]
3. Dr. Nathaniel Cady, Albany [Shewanella DOE work sparked collaboration on spatiotemporal simulations and nanotechnology experiments on microbial cross-feeding]
4. Dr. Christopher Marx, Harvard University [Multiple ongoing collaborations, including two recent DOE grants on simulation and experiments of microbial cross-feeding (also include Nate Cady, above)]
5. Lars Angenent: Department of Biological and Environmental Engineering, Cornell University, Ithaca, NY [collaboration of measuring gene expression in Shewanella during interaction with Lactococcus lactis]

Inventions/Patents, licensing agreements:
N/A

Theory behind Computer Modeling:
We summarize here the main method developed for extracting information about environemtnal triggers of gene expression. A more detailed description is available upon request. The full description of the method, as well as pointers to downloadable software are expected to be published in the near future.

We introduce a new method called Dynamic Detection of Transcriptional Triggers (D2T2), aimed at performing reverse engineering of responsive genes to external perturbations from gene expression time courses, through integration with steady-state profiles. The D2T2 algorithm requires a preliminary step, aimed at inferring a gene network based on steady state profiles. In this first step, a gene network model is trained on an independent large compendium of gene expression measurements taken under the assumption of steady state conditions, in analogy to previous reverse engineering approaches. The fundamental simplifying assumption of this model is that the rate of transcription of gene i (i.e. x_i) can be expressed as a linear combination of all other gene expression levels. Under steady state, this assumption reduces our model to a system of algebraic equations:

$$A_c x = 0$$

where A_c is a sparse square matrix (n×n, where n is the number of genes) and each non-zero element in Ac (i.e. a_{Cij}) represents the influence of gene *j* on gene *i*. This first preliminary step is based on a relevance network approach, where pairs of genes showing a high linear proportionality across multiple conditions are first selected in order to make the influence matrix sufficiently sparse. These relationships are further weighted through a multiple linear regression scheme, which evaluates the influence of a gene node in driving expression changes of the putative interacting ones. A gene-network model is then selected on the basis of a Bayesian criterion which identifies the best compromise between the model complexity and its predictive ability.

In a second step (the core of the D2T2 procedure), we go beyond the static influence matrix Ac computed above, and explicitly model dynamic changes induced upon perturbative events. This is achieved by modeling the rate of transcription of gene i (i.e. xi) as a linear combination of the "influencing-gene" expression levels plus an additional external input (i.e. u).

$$\dot{x} = Ax + Bu$$

Here b_i represents the influence that the perturbation u has in driving changes in the transcriptional rate of gene i. Hence, the condition-specific evolution in time of the system is described by a large-scale system of linear differential equations (ODEs) for the time-dependent gene expression profiles. Typically, the complexity of this type of model is limited by the sampling frequency and number of time points. Here, in order to overcome these limitations we take advantage of the initial steady state model in order to

reduce the fitting of the gene expression profiles to only two parameters (c and b, see Supplementary text of paper in preparation for full details):

$$\Delta x = CA_c x + B.$$

where Δx is the approximation of the continuous derivative in time using the Euler scheme and C is a diagonal matrix (nxn) consiting of normalizing factors c*ii*.

The identification of these two parameters for each single transcript aims at describing the portion of dynamical changes caused by external factors (i.e. B) from rearrangements induced by internal transcriptional regulatory mechanisms (Ac). In particular, the genes showing an incoherent behavior relative to the initial model (e.g. transient evolution) are chosen as the best candidates for direct targets of environmental changes. The underlying network and the significantly perturbed genes are identified through a rigorous statistical analysis. Once the distribution of empirical p-values had been generated, the q-value was used to correct the index of significance for multiple testing. Categories associated to a q-value in the 1% confidence interval were considered to be significant (similar results are obtained for the 5% confidence threshold).