Final Technical Report DE-FG02-04ER63942

Abstract

The transcription regulatory network is arguably the most important foundation of cellular function, since it exerts the most fundamental control over the abundance of virtually all of a cell's functional macromolecules. The two major components of a prokaryotic cell's transcription regulation network are the transcription factors (TFs) and the transcription factor binding sites (TFBS); these components are connected by the binding of TFs to their cognate TFBS under appropriate environmental conditions. Comparative genomics has proven to be a powerful bioinformatics method with which to study transcription regulation on a genome-wide level. We have further extended comparative genomics technologies that we introduced over the last several years. Specifically, we developed and applied statistical approaches to analysis of correlated sequence data (*i.e.,* sequences from closely related species). We also combined these technologies with functional genomic, proteomic and sequence data from multiple species, and developed computational technologies that provide inferences on the regulatory network connections, identifying the cognate transcription factor for predicted regulatory sites. Arguably the most important contribution of this work emerged in the course of the project. Specifically, the development of novel procedures of estimation and prediction in discrete high-D settings has broad implications for biology, genomics and well beyond. We showed that these procedures enjoy advantages over existing technologies in the identification of TBFS. These efforts are aimed toward identifying a cell's complete transcription regulatory network and underlying molecular mechanisms.

Summary

Accomplishments under this award fall into two general categories: bioinformatics technology development and bioinformatics applications to transcription regulation studies of bacterial species of environmental interest. In addition, we describe efforts toward resource sharing – providing open-source software as well as web services to the scientific community. Scientific publications resulting from this award are provided at the end of this report.

I. Bioinformatics technology development

A. Regulatory networks

We developed three regulatory network technologies. (i) We extended our previous Gibbs sampling models to incorporate phylogenetic relationships. We employed the Metropolis-Hasting sampling algorithm to draw samples from a full phylogenetic model and used centroid estimators for the cis-regulatory predictions[1]. We demonstrated that predictions using this model had improved sensitivity and positive predictive value over established methods, and showed that centroid estimators out-perform MAP estimators. (ii) We developed a technology (PhyloScan) for the identification of statistically significant matches to position weight matrices for related clades of species[2]. (iii) In collaboration with Gary Stormo, we developed a method to predict cognate transcription factors (TFs) for identified regulons, and thus extend regulatory network predictions to cis-trans connections[3].

B. Comparative genomics

We developed a sequence weighting procedure that minimizes the variances of parameter estimates[4]. We showed that, even with optimal weights, estimates of base frequency parameters

are inefficient for a clade of bacterial species, as well as a clade of mammalian species. We developed a method for estimating the effective sample size of sequences from clades of phylogenetically related species[5]. We developed tools to collect percent identity data for global alignments of orthologous sequences and generate summary statistics to assist the selection of appropriate species for motif finding studies[6].

C. Statistical inference in discrete high-D spaces

Arguably, the most important products of this research were the theoretical and methodological development of procedures for statistical inferences in discrete high-dimensional (high-D) spaces. First, we focused on the general problem of parameter estimation in discrete high-D spaces[7]. We applied statistical decision theory to show that there is no principled reason to expect the popular highest-scoring (HS) estimators, including MLEs, MAPs, maximum similarity, and minimum free energy, to be representative of the data's implication, and we found that the probability of an HS solution was often very small. To address these limitations, we developed alternative "centroid" estimators. We showed that centroids enjoy theoretical advantages over all of the HS estimators. Specifically, we showed that for binary and nominal variables, centroid estimators minimize expected pth power loss functions. In an important class of problems, centroids correspond to consensus estimators, and under squared-error loss the centroid is the feasible solution that is nearest to the mean. Accordingly, the centroid garners information from the entire ensemble of solutions to find an estimator that is representative of the entire posterior space. Secondly, we developed a general procedure for Bayesian confidence limits, a.k.a. credibility limits, of point estimates in discrete spaces, and illustrated its application to sequence alignment[8]. These confidence limits report on the global reliability of an alignment. Using promoters sequences from *Shewanella* species, we found that the reliability of alignments for ortholgous sequence pairs vary widely gene-to-gene and species-pair to species-pair. There is now clear evidence of the advantages of centroid estimators to predict ground truth standards in three important applications: the prediction of RNA secondary structures[9], protein structure prediction by homology[10], and motif finding[1]. Since in each case the probabilistic model of the centroid and the HS estimator are identical, these improvements in the prediction of ground truth reference sets stem entirely from differences in the estimation procedure.

II. Bioinformatics applications

A. *Rhodopseudomonas palustris*

We completed a genome-scale phylogenetic footprinting study focused on *R. palustris*[11]. This alpha-proteobacterial species carries out three of the chemical reactions that support life on this planet: the conversion of sunlight to chemical-potential energy, the conversion of carbon dioxide to cellular material, and the fixation of atmospheric nitrogen into ammonia. Our objective was to elucidate regulons in this bacterium using comparative data from 7 other alpha-proteobacterial species. Motifs were predicted upstream of 2,044 *R. palustris* genes and operons and clustered using the Bayesian Motif Clusterer (BMC[12]). Analysis of the resulting 101 motif clusters produced a number of significant findings, including: (i) the PpsR regulon, which controls the expression of many genes of the photosynthetic apparatus; (ii) the FlbD regulon, which controls flagellar synthesis; (iii) four nitrogen regulons (FixK2, NnrR, NtrC, Sigma54), representing an important first step in understanding nitrogen fixation in this species; (iv) a predicted cobalamine riboswitch; and (v) a organic hydroperoxide resistance regulon. Complete results are available at http://bayesweb.wadsworth.org/prokreg.html. We have also collaborated with the Center for Molecular and Cellular Systems (http://mippi.ornl.gov/) to identify motifs

upstream of genes encoding *R. palustris* proteins that interact to form complexes *in vivo*, thereby delineating the transcriptionally co-regulated genes (unpublished, with investigators of the CMCS). The PpsR motif was identified upstream of several photosynthesis and oxidation/reduction proteins that interact, and a novel motif was identified upstream of several interacting chemotaxis-related proteins.

B. *Shewanella oneidensis*

In collaboration with the Shewanella Federation, we examined the genome of *S. oneidensis* MR-1 for repetitive elements, transposons and pseudogenes, the delineation of which has improved the gene calls for hundreds of genes[13], and which will facilitate regulatory predictions by removing these repetitive sequences from the intergenic regions. These efforts, and the availability of several more *Shewanella* genomes, provide an exceptional data set for regulatory motif and regulon prediction. We have generated the orthologous promoter data sets of 17 *Shewanella* species and, using the Bayesian credibility limits[8], each data set is being analyzed to determine the appropriate clades of alignable sequences for analysis by the phylogenetic Gibbs sampler[1]. This work is on-going. Also, in collaboration with the Shewanella Federation, we have analyzed microarray expression data sets. Identification of the regulatory sites in co-expression data delineates the directly co-regulated genes (regulons) from those genes subject to secondary regulatory effects. In an analysis of 712 genes with altered expression profiles in a iron regulatory protein (Fur) mutant, we identified 56 genes that are likely directly regulated by Fur, by identifying the Fur regulatory motif in their upstream intergenic regions[14]. In addition, we identified 73 genes likely co-regulated by the regulatory protein EtrA, among 610 genes with altered expression in an EtrA mutant (unpublished, with J. Tiedje's lab).

III. Resource sharing and web site development supported by DE-FG02-04ER63942

All software developed and analyses of genomic data performed under this grant are available through web interfaces at http://ccmbweb.ccv.brown.edu/ and http://www.wadsworth.org/resnres/bioinfo/. Software tools include: Gibbs Sampler[15, 16], PhyloScan[2], and software for microbial species comparisons[6]. The websites allow the user to analyze sequence data, and provide access to user manuals, as well as online tutorials for phylogenetic footprinting and the analysis of prokaryotic co-expression data. Usage results are available for the Gibbs Sampler, which has processed over 4700 data sets. The open source Gibbs Sampler software is also distributed under GNU General Public License version 2 (http://www.gnu.org/copyleft/gpl.html). Since June 1, 2007, 1130 copies of the standalone versions of the Gibbs Sampler have been downloaded. This software has been used extensively in the identification of cis-regulatory signals[1, 11, 12, 14, 17-39].

IV. Students and Post-docs trained in part under funding from DE-FG02-04ER63942

Sean Conlan, Post-doc, Wadsworth Center. Dr. Conlan was a post-doctoral fellow at the Wadsworth Center with Drs. Lawrence and McCue prior to their moves (Dr. Lawrence to Brown University in mid-2004 and Dr. McCue to Pacific Northwest National Laboratory in late-2005). Dr. Conlan remained at the Wadsworth Center through early 2006, continuing to work remotely with Drs. Lawrence and McCue. Dr. Conlan was primarily responsible for our work with *Rhodopseudomonas palustris*[40], was instrumental to our work developing the phylogenetic Gibbs centroid sampler[1], and contributed significantly to several publications aimed at describing the use of our software tools[6, 16, 41].

Luis Carvalho, Graduate Student, Applied Math, Brown University. Mr. Carvalho is a Ph.D. student at Brown University, currently working with Dr. Lawrence on the properties of centroid estimators in discrete high-dimensional spaces[7].

Thomas Smith, Graduate Student, Computer Science, Rensselaer Polytechnic Institute. Dr. Smith was a graduate student at RPI (graduated spring 2005), where Dr. Lawrence had a joint appointment during his time at the Wadsworth Center. Dr. Smith contributed to the development of the phylogenetic Gibbs centroid sampler[1].

V. References

1. Newberg, L.A. et al. A phylogenetic Gibbs sampler that yields centroid solutions for cis regulatory site prediction. *Bioinformatics* **23**, 1718-1727 (2007).
2. Carmack, C.S., McCue, L.A., Newberg, L.A. & Lawrence, C.E. PhyloScan: identification of transcription factor binding sites using cross-species evidence. *Algorithms Mol Biol* **2**, 1 (2007).
3. Tan, K., McCue, L.A. & Stormo, G.D. Making connections between novel transcription factors and their DNA motifs. *Genome Res.* **15**, 312-320 (2005).
4. Newberg, L.A., McCue, L.A. & Lawrence, C.E. The Relative Inefficiency of Sequence Weights Approaches in Determining a Nucleotide Position Weight Matrix. *Statistical Applications in Genetics and Molecular Biology* **4**, 1-18 (2005).
5. Newberg, L.A. & Lawrence, C.E. Mammalian genomes ease location of human DNA functional segments but not their description. . *Statistical Applications in Genetics and Molecular Biology* **3**, 1-12 (2004).
6. Conlan, S. & McCue, L.A. Software to perform automated comparisons of pair-wise percent identities for microbial species. *Biotechniques* **40**, 578, 580,582 (2006).
7. Carvalho, L.E. & Lawrence, C.E. Centroid estimation in discrete high-dimensional spaces with applications in biology. *Proceedings of the National Academy of Sciences* **105**, 3209-3214 (2008).
8. Webb-Robertson, B.J., McCue, L.A. & Lawrence, C.E. Measuring Global Credibility with Application to Local Sequence Alignment. *PLoS Biol.* (2008).
9. Ding, Y.E., Chan, C.Y. & Lawrence, C.E. RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *RNA* **11**, 1157-1166 (2005).
10. Miyazawa, S. A reliable sequence alignment method based on probabilities of residue correspondences. *Protein Eng.* **8**, 999-1009 (1995).
11. Conlan, S., Lawrence, C. & McCue, L.A. Rhodopseudomonas palustris regulons detected by cross-species analysis of alphaproteobacterial genomes. *Appl Environ Microbiol* **71**, 7442-52 (2005).
12. Qin, Z.S. et al. Identification of co-regulated genes through Bayesian clustering of predicted regulatory binding sites. *Nature Biotechnology* **21**, 435-439 (2003).
13. Romine, M.F., Carlson, T.S., Norbeck, A.D., McCue, L.A. & Lipton, M.S. Identification of mobile elements and pseudogenes in the Shewanella oneidensis MR-1 genome. *Appl Environ Microbiol* **74**, 3257-65 (2008).
14. Wan, X.F. et al. Transcriptomic and proteomic characterization of the Fur modulon in the metal-reducing bacterium *Shewanella oneidensis*. *J Bacteriol* **186**, 8385-400 (2004).
15. Thompson, W., Rouchka, E.C. & Lawrence, C.E. Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucl. Acids. Res.* **31**, 3580-3585 (2003).

16. Thompson, W.A., Newberg, L.A., Conlan, S., McCue, L.A. & Lawrence, C.E. The Gibbs Centroid Sampler. *Nucl. Acids Res.* **35**, W232-W237 (2007).
17. McCue, L. et al. Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Res* **29**, 774-82 (2001).
18. McCue, L.A., Thompson, W., Carmack, C.S. & Lawrence, C.E. Factors influencing the identification of transcription factor binding sites by cross-species comparison. *Genome Res* **12**, 1523-32 (2002).
19. Thompson, W., Palumbo, M.J., Wasserman, W.W., Liu, J.S. & Lawrence, C.E. Decoding Human Regulatory Circuits. *Genome Res.* **14**, 1967-1974 (2004).
20. Florczyk, M.A. et al. A Family of acr-Coregulated Mycobacterium tuberculosis Genes Shares a Common DNA Motif and Requires Rv3133c (dosR or devR) for Expression. *Infect. Immun.* **71**, 5332-5343 (2003).
21. Sandelin, A. & Wasserman, W.W. Constrained Binding Site Diversity within Families of Transcription Factors Enhances Pattern Discovery Bioinformatics. *Journal of Molecular Biology* **338**, 207 (2004).
22. Michaloski, J.S., Galante, P.A.F. & Malnic, B. Identification of potential regulatory motifs in odorant receptor genes by analysis of promoter sequences. *Genome Res.* **16**, 1091-1098 (2006).
23. Levesque, M.P. et al. Whole-Genome Analysis of the SHORT-ROOT Developmental Pathway in Arabidopsis. *PLoS Biol.* **4**, e143 (2006).
24. Sandve, G.K. & F., D. A survey of motif discovery methods in an integrated framework. *Biology Direct* **1**, 1-11 (2006).
25. Salisbury, J., Hutchison, K.W. & Graber, J.H. A multispecies comparison of the metazoan 3'-processing downstream elements and the CstF-64 RNA recognition motif. *BMC Genomics* **7** (2006).
26. Zaslavsky, E. & Singh, M. A combinatorial optimization approach for diverse motif finding applications. *Algorithms Mol. Biol.* **1** (2006).
27. Eriksson, P.R. et al. Global Regulation by the Yeast Spt10 Protein Is Mediated through Chromatin Structure and the Histone Upstream Activating Sequence Elements. *Mol. Cell. Biol.* **25**, 9127-9137 (2005).
28. Marinescu, V., Kohane, I. & Riva, A. MAPPER: a search engine for the computational identification of putative transcription factor binding sites in multiple genomes. *BMC Bioinformatics* **6**, 79 (2005).
29. Alkema, W.B.L., Lenhard, B. & Wasserman, W.W. Regulog Analysis: Detection of Conserved Regulatory Networks Across Bacteria: Application to Staphylococcus aureus. *Genome Res.* **14**, 1362-1373 (2004).
30. Kechris, K., van Zwet, E., Bickel, P. & Eisen, M. Detecting DNA regulatory motifs by incorporating positional trends in information content. *Genome Biology* **5**, R50 (2004).
31. Hong, S.-J., Lessner, F.H., Mahen, E.M. & Keiler, K.C. Proteomic identification of tmRNA substrates. *Proceedings of the National Academy of Sciences* **104**, 17128-17133 (2007).
32. Glover, R.T., Kriakov, J., Garforth, S.J., Baughn, A.D. & Jacobs, W.R., Jr. The Two-Component Regulatory System senX3-regX3 Regulates Phosphate-Dependent Gene Expression in Mycobacterium smegmatis. *J. Bacteriol.* **189**, 5495-5503 (2007).
33. Gibb, E.A. & Edgell, D.R. Multiple Controls Regulate the Expression of mobE, an HNH Homing Endonuclease Gene Embedded within a Ribonucleotide Reductase Gene of Phage Aeh1. *J. Bacteriol.* **189**, 4648-4661 (2007).

34. Brouns, S.J.J. et al. Identification of the Missing Links in Prokaryotic Pentose Oxidation Pathways: evidence for enzyme recruitment. *J. Biol. Chem.* **281**, 27378-27388 (2006).
35. Clark, M.E. et al. Temporal Transcriptomic Analysis as Desulfovibrio vulgaris Hildenborough Transitions into Stationary Phase during Electron Donor Depletion. *Appl. Environ. Microbiol.* **72**, 5578-5588 (2006).
36. Erauso, G., Stedman, K.M., van de Werken, H.J.G., Zillig, W. & van der Oost, J. Two novel conjugative plasmids from a single strain of Sulfolobus. *Microbiology* **152**, 1951-1968 (2006).
37. Choi, K. & Kim, S. ComPath: comparative enzyme analysis and annotation in pathway/subsystem contexts. *BMC Bioinformatics* **9**, 145 (2008).
38. Kato, S., Kosaka, T. & Watanabe, K. Comparative transcriptome analysis of responses of Methanothermobacter thermautotrophicus to different environmental stimuli. *Environ Microbiol* **10**, 893-905 (2008).
39. da Rocha, R.P., de Miranda Paquola, A.C., do Valle Marques, M., Menck, C.F.M. & Galhardo, R.S. Characterization of the SOS Regulon of Caulobacter crescentus. *J. Bacteriol.* **190**, 1209-1218 (2008).
40. Conlan, S., Lawrence, C. & McCue, L.A. Rhodopseudomonas palustris Regulons Detected by Cross-Species Analysis of Alphaproteobacterial Genomes. *Appl. Environ. Microbiol.* **71**, 7442-7452 (2005).
41. Thompson, W., Conlan, S., McCue, L.A. & Lawrence, C.E. in Methods in Molecular Biology, Comparative Genomics (ed. Bergman, N.) 403-423 (Humana Press, 2007).

VI. Publications supported wholly or in part by DE-FG02-04ER63942

A. Research papers, technology development (in chronological order):

Mammalian genomes ease location of human DNA functional segments but not their description. Newberg LA and Lawrence CE. (2004) *Stat Appl Genet Mol Biol.* 3:Article 23.
http://dx.doi.org/10.2202/1544-6115.1065.
This paper analyzes the added statistical power that is obtainable from multi-species sequence data sets. It quantifies the extent to which additional genomes will be useful when attempting to locate and characterize transcription factor binding sites.

Making connections between novel transcription factors and their DNA motifs. Tan K, McCue LA, and Stormo GD. (2005) *Genome Res.* 15(2):312-20.
http://dx.doi.org/10.1101/gr.3069205.
This paper describes the development of computational methods to connect transcription factors and DNA motifs, using *Escherichia coli* as a model system. Our method uses three types of mutually independent information that are combined to calculate the probability of a given transcription-factor-DNA-motif pair being a true pair.

The relative inefficiency of sequence weights approaches in determining a nucleotide position weight matrix. Newberg LA, McCue LA, and Lawrence CE. (2005) *Stat Appl Genet Mol Biol.* 4:Article 13.
http://dx.doi.org/10.2202/1544-6115.1135.
This paper evaluates the use of sequence weights in inferences drawn from multiple evolutionarily related sequences.

Software to perform automated comparisons of pairwise percent identities for microbial species. Conlan S and McCue LA. (2006) *Biotechniques*. 40:578-582.
This paper describes, and provides a web address for easy download of, the tools that we use for calculating percent identities of promoter regions between species. These tools aid in the identification of species appropriate and useful for cross-species promoter analysis.

PhyloScan: Identification of transcription factor binding sites using cross-species evidence. Carmack CS, McCue LA, Newberg LA, and Lawrence CE. (2007) *Algorithms Mol Biol.* 2:1.
http://dx.doi.org/10.1186/1748-7188-2-1.
This paper describes the scanning algorithm, PhyloScan, that searches a genome-scale database with a position weight matrix by combining evidence from matching sites found in orthologous data from several related species with evidence from multiple sites within an intergenic region to increase the statistical power of regulon prediction.

A phylogenetic Gibbs sampler that yields centroid solutions for cis-regulatory site prediction. Newberg LA, Thompson WA, Conlan S, Smith TM, McCue LA, and Lawrence CE. (2007) *Bioinformatics*. 23(14):1718-1727.
http://dx.doi.org/10.1093/bioinformatics/btm241.
This paper describes a version of the Gibbs recursive sampler that incorporates the phylogeny of the input sequences through the use of an evolutionary model and calculates an ensemble centroid motif solution. Using simulated data, we show that false positive predictions, caused by correlation among the sequences, are dramatically reduced by these added features.

Centroid estimators for inference in high-dimensional discrete spaces. Carvalho LE and Lawrence CE. (2008) *Proc Natl Acad Sci U S A*, 105(9): 3209–3214.
http://dx.doi.org/10.1073/pnas.0712329105.
This paper reports on a novel procedure, centroid estimation, to obtain point estimates in discrete high-D spaces. Properties of these estimators are identified in four theorems, and evidence of improvements in the prediction of ground truth standards using these estimators compared to very popular highest scoring procedures is summarized.

Measuring global credibility with application to local sequence alignment. Webb-Robertson BJ, McCue LA, and Lawrence CE. (2008) *PLoS Comput Biol.* 4(5):e1000077.
http://dx.doi.org/10.1371/journal.pcbi.1000077.
This paper describes the development of Bayesian credibility limits to describe the uncertainty associated with high-dimensional inference problems, with a specific application to pairwise sequence alignment. We showed that credibility limits of the alignments of promoter sequences of 125 orthologous sequence pairs from six *Shewanella* species vary widely, and that centroid alignments dependably have tighter credibility limits than traditional maximum similarity alignments.

B. Research papers & book chapters, resource sharing (in chronological order):

Using the Gibbs Motif Sampler to find conserved domains in DNA and protein sequences. Thompson W, McCue LA, and Lawrence CE. (2005) in *Current Protocols in Bioinformatics*, (A.D. Baxevanis, D.B. Davison, R.D.M. Page, G.A. Petsko, L.D. Stein, and G.D. Stormo, eds.), John Wiley & Sons, Inc., New York, NY. pp. 2.8.1-2.8.38.
http://dx.doi.org/10.1002/0471250953.bi0208s10.

This book chapter describes the basic operation of the web interface to Gibbs and advanced examples of its use for locating transcription factor binding sites in unaligned DNA sequences.

Using the Gibbs Motif Sampler for Phylogenetic Footprinting. Thompson W, Conlan S, McCue LA, and Lawrence CE. (2007) in *Methods in Molecular Biology, Comparative Genomics* (N. Bergman ed.), Humana Press, Inc., Totowa, NJ. vol. 395, pp. 403-424.
This book chapter describes the use of the Gibbs Recursive Sampler to locate transcription factor binding sites in a collection of orthologous nucleotide sequences, *i.e.* phylogenetic footprinting.

The Gibbs Centroid Sampler. Thompson WA, Newberg LA, Conlan S, McCue LA, and Lawrence CE. (2007) *Nucleic Acids Res.* 35:W232-237.
http://dx.doi.org/10.1093/nar/gkm265.
This paper, in the Web Server issue of *Nucleic Acids Research*, describes the advanced features of the web interface to Gibbs that allow identification of the centroid solution.

C. Research papers, biology applications (in chronological order):

Transcriptomic and proteomic characterization of the Fur modulon in the metal-reducing bacterium *Shewanella oneidensis*. Wan XF, Verberkmoes NC, McCue LA, Stanek D, Connelly H, Hauser LJ, Wu L, Liu X, Yan T, Leaphart A, Hettich RL, Zhou J, and Thompson DK. (2004) *J Bacteriol.* 186(24):8385-400.
http://dx.doi.org/10.1128/JB.186.24.8385-8400.2004.
This paper is the result of collaboration with members of the Shewanella Federation to examine regulation by the transcription factor Fur in *Shewanella oneidensis* MR-1 by integrating DNA microarrays, proteomics, and promoter sequence analysis.

Global profiling of *Shewanella oneidensis* MR-1: expression of hypothetical genes and improved functional annotations. Kolker E, Picone AF, Galperin MY, Romine MF, Higdon R, Makarova KS, Kolker N, Anderson GA, Qiu X, Auberry KJ, Babnigg G, Beliaev AS, Edlefsen P, Elias DA, Gorby YA, Holzman T, Klappenbach JA, Konstantinidis KT, Land ML, Lipton MS, McCue LA, Monroe M, Pasa-Tolic L, Pinchuk G, Purvine S, Serres MH, Tsapin S, Zakrajsek BA, Zhu W, Zhou J, Larimer FW, Lawrence CE, Riley M, Collart FR, Yates JR 3rd, Smith RD, Giometti CS, Nealson KH, Fredrickson JK, and Tiedje JM. (2005) *Proc Natl Acad Sci U S A.* 102(6):2099-104.
http://dx.doi.org/10.1073/pnas.0409111102.
This paper was a joint effort by many members of the Shewanella Federation. The focus of this report was to annotate hypothetical ORFs in the *S. oneidensis* MR-1 genome.

***Rhodopseudomonas palustris* regulons detected by cross-species analysis of alpha-proteobacterial genomes.** Conlan S, Lawrence C, and McCue LA. (2005) *Appl. Environ. Microbiol.* 71:7442-7452.
http://dx.doi.org/10.1128/AEM.71.11.7442-7452.2005.
This paper describes a phylogenetic footprinting study focused on the identification of regulons in *R. palustris*. A total of 4,963 regulatory motifs were predicted and clustered into 101 putative regulons.

Identification of mobile elements and pseudogenes in the *Shewanella oneidensis* MR-1 genome. Romine MF, Carlson TS, Norbeck AD, McCue LA, and Lipton MS. (2008) *Appl. Environ. Microbiol.* 74(10):3257-3265.
http://dx.doi.org/10.1128/AEM.02720-07.
This paper is the result of collaboration with members of the Shewanella Federation to examine the *Shewanella oneidensis* MR-1 genome for repetitive elements, transposons, and pseudogenes.