Final Summary of "Interdisciplinary Study of *Shewanella oneidensis* MR-1's Metabolism & Metal Reduction"

This report describes the study supported by the Department of Energy (DOE)'s *Genomes to Life* grant DE-FG02-03ER63527 to PI Dr. E. Kolker at The BIATECH Institute (formerly BIATECH).

Introduction

The *Shewanella Federation* is a multi-investigator and cross-institutional consortium formed to characterize and model the biology of the metabolically versatile bacterium *Shewanella oneidensis* MR-1. This federation brings together the diversity of scientific expertise necessary to understand the biology of a microorganism at a whole-system level. Information gathered from these systems-level investigations is aimed at modeling cellular networks to construct a computation model, predicting the biology of *S. oneidensis* MR-1 in general and its metabolic and metal-reduction capabilities in particular. The *Shewanella Federation* is one of the leading consortia within the DOE's Genomes-To-Life (GTL) Initiative.

As an integral part of the *Shewanella Federation*, our project focused primarily on analysis of different types of data produced by global high-throughput technologies, data integration of gene annotation, and gene and protein expression information, as well as on getting a better functional annotation of *Shewanella* genes. Specifically, four of our numerous major activities and achievements include the development of:

1. **New** statistical models for identification and expression proteomics, superior to currently available approaches (including our own earlier ones);
2. **New** approaches to improve gene annotations on the whole-organism scale;
3. **New** standards for annotation, transcriptomics and proteomics approaches;
4. **New** generalized approaches for data integration of gene annotation, gene and protein expression information.

Due to the space limitations, we describe in detail points 1 and 2 and briefly summarize 3 and 4.

Summary of Accomplishments

1. Statistical Models for Proteomics

One of our principal achievements in proteomics analysis was the ability to effectively identify proteins in MS/MS analysis only identified by a single unique peptide; so called "one-hit-wonders" (Higdon et *al* 2007). To avoid false-positive protein identification, 2 or more unique peptides identified within a single protein are generally recommended. Still, in a typical high-throughput experiment, hundreds of proteins are identified only by a single peptide. We introduced a unique method for distinguishing between true and false identifications among single-hit proteins. The approach is based on randomized database searching and usage of logistic regression models with cross-validation. This approach was implemented to analyze three bacterial samples, developed by several *Shewanella Federation* laboratories, and primarily based on the work by the Pacific Northwest National Laboratory and our own lab. Our original approach enables recovery 68–98% of the correct single-hit proteins with an error rate of <2%. This results in a 22–65% increase in number of identified proteins. Identifying true single-hit proteins will lead to discovering many crucial regulators, biomarkers and other low-abundance proteins.

Other advancements in proteomics analysis led to the development of our logistic identification of peptide sequences (LIPS) models for identifying peptides (Higdon et *al* 2004). These are statistically valid, customizable, and easy-to-use models for identifying peptides in high-throughput proteomics experiments. We have enhanced the use of randomized database searches to estimate false discovery rates for peptide and protein identifications (Higdon *et al* 2005). This approach has led to experiment-specific estimates of peptide identification probabilities to be used as thresholds for identification, and as a basis for estimating the false discovery rate for protein identification.

2. Improving Genome Annotation

One the most novel and significant achievements of this project was the collaborative integrative multi-institutional study to annotate hypothetical genes, led by our lab and published in *PNAS* (Kolker *et al* 2005). Similar to most other sequenced organisms, approximately 40% of the predicted ORFs in the *S. oneidensis* genome were

annotated as uncharacterized hypothetical genes. We implemented an original integrative approach by using experimental and computational analyses to provide more detailed insight into gene function. Transcriptomic and proteomic analyses confidently identified 538 hypothetical genes as expressed in *S. oneidensis* cells both as mRNAs and proteins (33% of all predicted hypothetical proteins). Publicly available analysis tools as well as databases and the expression data were applied to improve the annotation of these genes.

The annotation results were scored by using a unique seven-category schema that ranked both confidence and precision of the functional assignment. We were able to identify homologs for nearly all of these hypothetical proteins (97%), but could confidently assign exact biochemical functions for only 16 proteins (category 1; 3%). Altogether, computational and experimental evidence provided functional assignments or insights for 240 more genes (categories 2–5; 45%). These functional annotations advance our understanding of genes involved in vital cellular processes, including energy conversion, ion transport, secondary metabolism, and signal transduction. We showed that this integrative approach offers a valuable means to undertake the enormous challenge of characterizing the rapidly growing number of hypothetical proteins with each newly sequenced genome.

3. New standards

All the above advancements were possible due to new standards developed for annotation, transcriptomics and proteomics approaches. New standards include both computational and experimental work, and were described in detail in 8 manuscripts, including the two most recently published in *Nature Biotechnology* (Nos. 24 and 25). Our work on the proteomics standards is currently supported by the NIH grant.

4. Data Integration

A fourth key area of development was in the integration of experimental high-throughput proteomics and genomics data with genome annotation data. This has resulted in a development of new integrative approaches in a general and new database system on *S. oneidensis,* called SODB. These approaches have been further generalized for use with multiple organisms and are currently supported by the NSF grant.

Conclusions

Analysis of high-throughput data, the annotation of hypothetical proteins and the integration of this data are key to the success of modern biology. This project has made advances by overcoming bottlenecks in these areas that are crucial to understanding the biology of organisms such as *S. oneidensis,* and by enhancing our ability to deduce the complex networks and pathways responsible for the function and versatility of *S. oneidensis* and other organisms.

References:

1. Goodlett, D. R.; Keller, A.; Watts, J. D.; Newitt, R.; Yi, E. C.; Purvine, S.; Eng, J. K.; von Haller, P.; Aebersold, R.; Kolker, E., Differential stable isotope labeling of peptides for quantitation and de novo sequence derivation. *Rapid Commun Mass Spectrom* 2001, 15, (14), 1214-21.
2. Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R., Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem* 2002, 74, (20), 5383-92.
3. Keller, A.; Purvine, S.; Nesvizhskii, A. I.; Stolyar, S.; Goodlett, D. R.; Kolker, E., Experimental protein mixture for validating tandem mass spectral analysis. *Omics* 2002, 6, (2), 207-12.
4. Kolker, E.; Purvine, S.; Picone, A.; Cherny, T.; Akerley, B. J.; Munson, R. S., Jr.; Palsson, B. O.; Daines, D. A.; Smith, A. L., H. influenzae Consortium: integrative study of H. influenzae-human interactions. *Omics* 2002, 6, (4), 341-8.
5. Tjaden, B.; Saxena, R. M.; Stolyar, S.; Haynor, D. R.; Kolker, E.; Rosenow, C., Transcriptome analysis of Escherichia coli using high-density oligonucleotide probe arrays. *Nucleic Acids Res* 2002, 30, (17), 3732-8.
6. Tjaden, B.; Haynor, D. R.; Stolyar, S.; Rosenow, C.; Kolker, E., Identifying operons and untranslated regions of transcripts using Escherichia coli RNA expression analysis. *Bioinformatics* 2002, 18 Suppl 1, S337-44.
7. Kolker, E.; Purvine, S.; Galperin, M. Y.; Stolyar, S.; Goodlett, D. R.; Nesvizhskii, A. I.; Keller, A.; Xie, T.; Eng, J. K.; Yi, E.; Hood, L.; Picone, A. F.; Cherny, T.; Tjaden, B. C.; Siegel, A. F.; Reilly, T. J.; Makarova, K. S.; Palsson, B. O.; Smith, A. L., Initial proteome analysis of model microorganism Haemophilus influenzae strain Rd KW20. *J Bacteriol* 2003, 185, (15), 4593-602.
8. Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R., A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem* 2003, 75, (17), 4646-58.
9. Higdon, R.; Kolker, N.; Picone, A.; van Belle, G.; Kolker, E., LIP index for peptide classification using MS/MS and SEQUEST search via logistic regression. *Omics* 2004, 8, (4), 357-69.
10. Holzman, T.; Kolker, E., Statistical analysis of global gene expression data: some practical considerations. *Curr Opin Biotechnol* 2004, 15, (1), 52-7.

11. Kolker, E. S., Framing as a cultural resource in health social movements: funding activism and the breast cancer movement in the US 1990-1993. *Sociol Health Illn* 2004, 26, (6), 820-44.

12. Kolker, E.; Makarova, K. S.; Shabalina, S.; Picone, A. F.; Purvine, S.; Holzman, T.; Cherny, T.; Armbruster, D.; Munson, R. S., Jr.; Kolesov, G.; Frishman, D.; Galperin, M. Y., Identification and functional analysis of 'hypothetical' genes expressed in Haemophilus influenzae. *Nucleic Acids Res* 2004, 32, (8), 2353-61.

13. Purvine, S.; Kolker, N.; Kolker, E., Spectral quality assessment for high-throughput tandem mass spectrometry proteomics. *Omics* 2004, 8, (3), 255-65.

14. Purvine, S.; Picone, A. F.; Kolker, E., Standard mixtures for proteome studies. *Omics* 2004, 8, (1), 79-92.

15. Raghunathan, A.; Price, N. D.; Galperin, M. Y.; Makarova, K. S.; Purvine, S.; Picone, A. F.; Cherny, T.; Xie, T.; Reilly, T. J.; Munson, R., Jr.; Tyler, R. E.; Akerley, B. J.; Smith, A. L.; Palsson, B. O.; Kolker, E., In Silico Metabolic Model and Protein Expression of Haemophilus influenzae Strain Rd KW20 in Rich Medium. *Omics* 2004, 8, (1), 25-41.

16. Higdon, R.; Hogan, J. M.; Van Belle, G.; Kolker, E., Randomized sequence databases for tandem mass spectrometry peptide and protein identification. *Omics* 2005, 9, (4), 364-79.

17. Hogan, J. M.; Higdon, R.; Kolker, N.; Kolker, E., Charge state estimation for tandem mass spectrometry proteomics. *Omics* 2005, 9, (3), 233-50.

18. Kolker, E.; Picone, A. F.; Galperin, M. Y.; Romine, M. F.; Higdon, R.; Makarova, K. S.; Kolker, N.; Anderson, G. A.; Qiu, X.; Auberry, K. J.; Babnigg, G.; Beliaev, A. S.; Edlefsen, P.; Elias, D. A.; Gorby, Y. A.; Holzman, T.; Klappenbach, J. A.; Konstantinidis, K. T.; Land, M. L.; Lipton, M. S.; McCue, L. A.; Monroe, M.; Pasa-Tolic, L.; Pinchuk, G.; Purvine, S.; Serres, M. H.; Tsapin, S.; Zakrajsek, B. A.; Zhu, W.; Zhou, J.; Larimer, F. W.; Lawrence, C. E.; Riley, M.; Collart, F. R.; Yates, J. R., 3rd; Smith, R. D.; Giometti, C. S.; Nealson, K. H.; Fredrickson, J. K.; Tiedje, J. M., Global profiling of Shewanella oneidensis MR-1: expression of hypothetical genes and improved functional annotations. *Proc Natl Acad Sci U S A* 2005, 102, (6), 2099-104.

19. Louie, B.; Mork, P.; Shaker, R.; Kolker, N.; Kolker, E.; Tarczy-Hornoch, P., Integration of data for gene annotation using the BioMediator system. *AMIA Annu Symp Proc* 2005, 1036.

20. Galperin, M. Y.; Kolker, E., New metrics for comparative genomics. *Curr Opin Biotechnol* 2006, 17, (5), 440-7.

21. Hogan, J. M.; Higdon, R.; Kolker, E., Experimental standards for high-throughput proteomics. *Omics* 2006, 10, (2), 152-7.

22. Kolker, E.; Higdon, R.; Hogan, J. M., Protein identification and expression analysis using mass spectrometry. *Trends Microbiol* 2006, 14, (5), 229-35.

23. Higdon, R.; Kolker, E., A predictive model for identifying proteins by a single peptide match. *Bioinformatics* 2007, 23, (3), 277-80.

24. Field D, Garrity G, Selengut J, … Kolker E et al, Towards richer descriptions of our collection of genomes and metagenomes: the 'Minimum Information about a Genome Sequence' (MIGS) specification, *Nature Biotechnology*, 2007, in press.

25. Taylor C, Field D, Sansone A, … Kolker E et al, Promoting Coherent Minimum Reporting Requirements for Biological and Biomedical Investigations: The MIBBI Project, *Nature Biotechnology*, 2007, in press.