Milestone 1

***The Code*—Understand Gene Structure and Functional Potential of Plants and Microbes and Their Communities**

Section 1

Organism Sequencing, Annotation, and Comparative Genomics

1

U.S. DOE Joint Genome Institute Microbial Sequencing: Genomes to Life Projects

David Bruce[1]* (dbruce@lanl.gov), Tom Brettin[1], Patrick Chain[3], Cliff Han[1], Loren Hauser[5], Nikos Kyrpides[2], Miriam Land[5], Alla Lapidus[2], Frank Larimer[5], Jeremy Schmutz[4], Paul Gilna[1], Eddy Rubin[2], and Paul Richardson[2]

[1]JGI-Los Alamos National Laboratory, Los Alamos, NM; [2]JGI-Production Genomics Facility and Lawrence Berkeley National Laboratory, Berkeley, CA; [3]JGI-Lawrence Livermore National Laboratory, Livermore, CA; [4]JGI-Stanford Human Genome Center, Palo Alto, CA; and [5]JGI-Oak Ridge National Laboratory, Oak Ridge, TN

The US DOE Joint Genome Institute (JGI) sequences microbial and metagenomic projects through three main programs: DOE Microbial Genome Program (MGP), JGI Community Sequencing Program (CSP) and DOE Genomes to Life Program (GTL). The principle goal of the MGP is to fund sequencing projects related to DOE interests, the principle goal of the CSP is to fund sequencing projects from a broad range of disciplines that may not be covered in the MGP, and the principle goal of the GTL sequencing projects is to fund sequencing projects in direct support of the GTL program. The JGI is responsible for sequencing, assembling, annotating microbial genomes, and publishing sequence and annotation in GenBank and the DOE JGI Integrated Microbial Genomics web based system. The JGI has sequenced nearly 250 microbes and metagenomic samples to draft quality and completely finished over 120 microbes. Most microbial projects are targeted for finishing. The overall capacity is now approximately 100-125 microbial projects per year through draft sequencing and finishing. Virtually all microbial projects are sequenced by the whole genome shotgun method. To being the sequencing process, the Library group randomly shears the purified DNA under different conditions and selects for three size populations. Fragments are end repaired and selected for inserts in the range of 3kb, 8kb, and 40kb. These are cloned into different vector systems and checked for quality by PCR or sequencing. The libraries are sequenced by the Production group to approximately 8.5X coverage. The resulting reads are trimmed for vector sequences and assembled. The assembly is quality checked, automatically annotated by the Annotation group, and released to the collaborating PI as the initial Quality Draft assembly. For finishing, the draft assembly is assigned to a Finishing group. The Finishing group closes all sequence gaps, resolves all repeat discrepancies, and improves all low quality regions. The final assembly is then passed to the Quality Assurance group to assess the integrity and overall quality of the genome sequence. The finished sequence then receives a final annotation and this package is used as the basis for analysis and publication in GenBank and the DOE JGI Integrated Microbial Genomics web based system. The JGI is made up of affiliates from a number of national laboratories including Lawrence Berkeley National Laboratory, Lawrence Livermore National Laboratory, Los Alamos National Laboratory, Oak Ridge National Laboratory, and the Stanford Human Genome Center.

* Presenting author

2

High Throughput Genome Annotation for U.S. DOE Joint Genome Institute Microbial Genomes

Miriam Land* (landml@ornl.gov), Loren Hauser, Phil LoCascio, Gwo-Liang Chen, Denise Schmoyer, and Frank Larimer

Oak Ridge National Laboratory, Oak Ridge TN

http://genome.ornl.gov/microbial/

The U.S. DOE Joint Genome Institute (JGI) performs high-throughput sequencing and annotation of microbial genomes through the DOE Microbial Genome Program (MGP). The world-wide rate of sequencing is resulting in a rapid expansion of microbial genomic data, which requires the development of comprehensive automated tools to provide in-depth annotation which can keep pace with the expanding microbial dataset. We have and continue to develop tools for genome analysis that provide automated, regularly updated, comprehensive annotation of microbial genomes using consistent methodology for gene calling and feature recognition. We have developed and continue to improve a genome annotation pipeline. The pipeline includes gene calls, multiple database searches, prediction of RNAs, and other annotation tools as they become available for a diverse and automated annotation.

Comprehensive representation of microbial genomes requires deeper annotation of structural features, including operon and regulon organization, promoter and ribosome binding site recognition, miscellaneous RNAs, and other functional elements. Linkage and integration of the gene/protein/function catalog to phylogenomic, structural, proteomic, transcriptional, and metabolic profiles are being developed. The expanding set of microbial genomes comprises an extensive resource for comparative genomes: new tools continue to be developed for rapid exploration of gene and operon phylogeny, regulatory networking, and functional proteomics.

A major continuing activity involves the public release of the data. Each genome is supported with a web site of the automated annotation, the data are submitted to GenBank for broader release and the data are prepared for the JGI's Integrated Microbial Genomes (IMG) database. The IMG resource is updated quarterly, in addition to the continuous addition of new genomes from JGI. 50-100 new projects will be initiated annually by JGI that require annotation. The deep sequencing of specific genera as well as specialized (physiological and phylogenetic) groups requires new views and analytical schemes.

The JGI is made up of affiliates from a number of national laboratories including Lawrence Berkeley National Laboratory, Lawrence Livermore National Laboratory, Los Alamos National Laboratory, Oak Ridge National Laboratory, and the Stanford Human Genome Center.

* Presenting author

3

Understanding Microbial Genomic Structures and Applications to Biological Pathway Inference

Z. Su[1], F. Mao[1], H. Wu[1], P. Dam[1], X. Chen[2], T. Jiang[2], V. Olman[1], B. Palenik[3], and **Ying Xu**[1]* (xyn@bmb.uga.edu)

[1]University of Georgia, Athens, GA; [2]University of California, Riverside, CA; and [3]Scripps Institution of Oceanography, University of California, San Diego, CA

The rapid increase in the number of sequenced microbial genomes provides unprecedented opportunities to computational biologists to decipher the genomic structures of these microbes through development and application of advanced comparative genome analysis tools. In this presentation, we describe a systematic study we have been carrying out on deciphering microbial genomic structures and linking the discovered genomic structures to prediction of metabolic pathways. This study consists of the following three main components: (a) deciphering microbial genomic structures and discovering new ones through development and application of advanced comparative genome analysis tools, (b) systematic study of relationships between microbial genomic structures and metabolic pathways through mapping all KEGG pathways to over 300 microbial genomes, and (c) application of the discovered relationships between genomic structures and pathways to prediction of biological pathways and networks.

A. Deciphering microbial genomic structures

We have recently developed a computer program JPOP[1,2] for operon structure predictions in both prokaryotic and archaea genomes. Testing on *E. coli.* data with experimental validation indicates that the program has an prediction accuracy about 80%. Since the publication of JPOP, a couple of operon prediction programs have been published including VIMSS[10] and Pathway Tools[11], reaching similar levels of prediction accuracy. Using these programs, we have made operon prediction for 300+ microbial genomes (all data are available upon request). This data set not only provides a rich source of information for our prediction of biological pathways and networks (see section C), but also facilitates investigation of higher level and less understood structures in microbial genomes. Through comparative genome analyses of 300+ microbial genomes, we have recently firmly established uber-operon, a concept introduced a few years ago by other authors, as a layer of genomic structures, which have direct implications to biological pathway predictions[3]. For example, we have demonstrated that a number of well studied metabolic pathways are made of (genes of) a small number of uber-operons (*versus* a large number of operons) [3]. In addition, we have established some interesting relationships between uber-operons and regulons, which have established a solid stepping stone for us to develop a computer program for regulon prediction in general *via* prediction of uber-operons. We have also recently developed an effective paradigm for predicting *cis* regulatory elements[4], through comparative analysis of closed related genomes, providing another important piece of information for regulon prediction. We expect that we will be able to develop the first computer program for regulon prediction in the very near future.

B. Systematic mapping of metabolic pathways to microbial genomes

The metabolic pathways of KEGG database provides a rich source of information, which can be directly mapped to individual genomes. However until very recently, there has not been an effective way for mapping KEGG pathways to genomes other than the simple minded approach through sequence similarity search. We have recently demonstrated that BLAST search or its variations/

* Presenting author

generalizations such as bi-direction best hit (BDBH) or COG search do not provide satisfactory mapping results[5] as virtually all these methods attempt to find orthologous gene relationship using sequence similarity information alone. We have recently developed a computer program P-MAP for mapping orthologous genes in the context of pathway mapping using both sequence similarity information and genomic structure information, having substantially improved the mapping accuracy of pathways. The basic idea of P-MAP pathway mapping is that it attempts to map genes of a pathway to their homologous genes in the target genome, under the condition that these mapped genes are grouped into a (small) number of operons. The limitation of the current P-MAP algorithm is that it assumes that a template pathway is given in a form that its individual components have genes assigned in the template genome, limiting direct applications of KEGG (template) pathways. We have recently generalized the framework of P-MAP, allowing mapping a generic pathway model (consisting of enzymes and enzymatic reasons rather than specific genes assigned to each enzyme) to a target genome, by mapping individual enzymes to genes that are grouped into a number of operons in the target genome[6]. Using this novel capability, we have mapped metabolic pathways of KEGG to 300+ microbial genomes (data are available upon request). A detailed analysis is currently under way, attempting to understand the general relationship between metabolic pathways and operon, uber-operon and regulon structures. We expect that this analysis will lead to new understanding about genomic structures, the organization and evolution of metabolic pathways, which is expected to be done within the next few weeks.

C. Pathway predictions through application of identified genomic structures:

As we understand now, genomic structures such as operons, uber-oprons and regulons and their detailed organizations provide significant amount of information about the component genes and even wiring diagrams of metabolic pathways. Using such information derived through prediction of operons, uber-operons and regulons, we have made a number of predictions of non-trivial biological networks, including phosphorus assimilation pathways[7], carbon fixation pathways[8], and nitrogen assimilation pathways[9] in *cyanobacteria* and a cross-talk network between nitrogen assimilation and photosynthesis[4], for which we for the first time proposed a detailed molecular mechanism how these two processes orchestrate with each other. These predictions have provided a number of new insights about these important biological processes. By extending these predictions, we are currently focused on prediction of a group of pathways relevant to carbon sequestration. Detailed results of these predicted pathway presentation will be presented at the GTL workshop.

Acknowledgement: This project is supported by the U.S. Department of Energy's Genomics:GTL Program under project "Carbon Sequestration in *Synechococcus sp*: From Molecular Machines to Hierarchical Modeling" (http://www.genomes-to-life.org).

References

1. Chen, X., Su, Z., Dam, P., Palenik, B., Xu, Y. and Jiang, T. (2004) "Operon prediction by comparative genomics: an application to the *Synechococcus sp.* WH8102 genome," *Nucleic Acids Res*, 32, 2147-2157.
2. Chen, X., Su, Z., Xu, Y. and Jiang, T. (2004) "Computational Prediction of Operons in Synechococcus sp. WH8102.,"*Genome Inform Ser Workshop Genome Inform*, 15, 211-222 (best paper award).
3. Che, G. Li, F. Mao, H. Wu, and Ying Xu, "Detecting uber-operons in microbial genomes," submitted to *Proc Natl Acad Sci USA*, 2005.
4. Z. Su, F. Mao, V. Olman and Ying Xu, "Comparative genomics analyses of ntcA regulons in cyanobacteria: regulation of nitrogen assimilation and its coupling to photosynthesis," *Nucleic Acids Research*, vol 33(16): 5156 - 5171, 2005.
5. Mao, Z. Su, V. Olman, P. Dam, Z. Liu, Ying Xu, "Mapping of orthologous genes in the context of biological pathways: an application of integer programming," *Proc Natl Acad Sci USA*, 2005 (in press).

* Presenting author

6. Mao, W. Wu, Ying Xu, "Mapping of KEGG metabolic pathways to microbial genomes", submitted, 2005.

7. Z. Su, A. Dam, X. Chen, V Olman, T. Jiang, B. Palenik, and Ying Xu, Computational Inference of Regulatory Pathways in Microbes: an application to the construction of phosphorus assimilation pathways in Synechococcus WH8102, *Genome Informatics* pp. 3 - 13, vol 14, Universal Academy Publishing, 2003.

8. P. Dam, Z. Su, V Olman, Ying Xu, *In silico* construction of the carbon fixation pathway in Synechococcus sp. WH8102, *Journal of Biological Systems*, vol. 12, pp.97-125, 2004.

9. Z. Su, P. Dam, F. Mao, V. Olman, I. Paulsen, B. Palenik and Ying Xu, Computational inference and experimental validation of nitrogen assimilation regulatory networks in cyanobacterium Synechococcus sp. WH8102, *Nucleic Acids Research* , 2005 (in press).

10. Price, M.N., et al., *A novel method for accurate operon predictions in all sequenced prokaryotes. Nucleic Acids Res*, 2005. 33(3): p. 880-92.

11. Karp, P.D., S. Paley, and P. Romero, *The Pathway Tools software.* Bioinformatics, 2002. 18 Suppl 1: p. S225-S232.

4 MEWG

The BioCyc Collection of 200 Pathway/Genome Databases and the MetaCyc Database of Metabolic Pathways and Enzymes

Peter D. Karp[1]* (pkarp@ai.sri.com), Christos Ouzounis[2], and Sue Rhee[3]

[1]SRI International, Menlo Park, CA; [2]European Bioinformatics Institute, Hinxton, UK; and [3]Carnegie Institution, Stanford, CA

The BioCyc Database Collection[1] is a set of 200 Pathway/Genome Databases (PGDBs) for most prokaryotic and eukaryotic organisms whose genomes have been completely sequenced to date. The BioCyc collection provides a unique resource for metabolic engineering and for global and comparative analyses of genomes and metabolic networks.

Each organism-specific PGDB within BioCyc contains the complete genome of the organism plus the following additional information inferred by the Pathway Tools[2] software:

- Predicted metabolic pathways as inferred from the MetaCyc[3] database
- Predicted genes to fill holes in the metabolic pathways (pathway holes are pathway steps for which no enzyme has been identified in the genome)
- Predicted operons for each bacterial PGDB
- Transport reactions inferred from the product descriptions of transport proteins by the Transport Inference Parser
- A metabolic overview diagram containing the metabolic enzymes, transport proteins, and membrane proteins of each organism is constructed automatically

The BioCyc collection can be accessed in several ways including interactive access via the BioCyc.org web site, bulk downloading in several formats including Systems Biology Markup Language (SBML) and BioPAX, and querying within SRI's BioWarehouse system for database integration. Most BioCyc PGDBs are freely and openly available to all.

* Presenting author

We seek scientists to adopt and curate individual PGDBs within the BioCyc collection. Only by harnessing the expertise of many scientists can we hope to produce biological databases that accurately capture the depth and breadth of biomedical knowledge. To adopt a database, send email to biocyc-support@ai.sri.com.

The Pathway Tools software that powers the BioCyc Web site provides powerful query and visualization operations for each BioCyc database. For example, the Omics viewer allows scientists to visualize combinations of gene expression, proteomics, and metabolomics data on the metabolic map of an organism (see http://biocyc.org/ov-expr.shtml). A genome browser permits interactive exploration of either a single genome, or of orthologous regions of multiple genomes. A newly developed set of comparative genomics tools supports many comparisons across the genomes and metabolic networks of the BioCyc collection. See http://biocyc.org/samples.shtml for an overview of BioCyc Web site functionality.

The MetaCyc database[3] describes experimentally elucidated metabolic pathways and enzymes as reported in the experimental literature. MetaCyc is both an online reference source on metabolic pathways and enzymes, and a solid foundation of experimentally proven pathways for use in computational pathway prediction. MetaCyc version 9.6 describes 690 pathways from more than 600 organisms. The 5500 biochemical reactions in MetaCyc reference 4800 chemical substrates, most of which contain chemical structure information. MetaCyc describes the properties of 3000 enzymes, such as their subunit structure, cofactors, activators, inhibitors, and in some cases their kinetic parameters. The information in MetaCyc was obtained from more than 8500 research articles, and emphasizes pathways and enzymes from microbes and plants.

References

1. P.D. Karp et al, "Expansion of the BioCyc Collection of Pathway/Genome Databases to 160 Genomes," *Nucleic Acids Research* 33:6083-9 2005.
2. P.D. Karp et al, "The Pathway Tools Software," *Bioinformatics* 18:S225-32 2002.
3. R. Caspi et al, "MetaCyc: A multiorganism database of metabolic pathways and enzymes," *Nucleic Acids Research* in press, 2006 Database issue.

5

Bioinformatic Methods Applied to Prediction of Bacterial Gene Function

Michelle Green[1]* (green@ai.sri.com), Balaji Srinivasan[2], Peter Karp[1], and **Harley McAdams**[2]

[1]SRI International, Menlo Park, CA and [2]Stanford University, Stanford, CA

We have applied two different bioinformatic methods to prediction of the function of *Caulobacter* gene products. First, we used the PathoLogic program to construct Pathway/Genome databases using the genome's annotation to predict the set of metabolic pathways present. PathoLogic determines the set of reactions composing those pathways from the list of enzymes in the organism. Enzymes in a genome are often missed or assigned a non-specific function (e.g., "thiolase family protein") during the initial annotation. These incomplete annotations result in "pathway holes" where the genome appears to lack the enzymes known to be needed to catalyze reactions in a pathway. Second, we combined gene co-conservation determined from 230 sequenced bacterial genomes with four

* Presenting author

types of functional genomic data to predict protein interaction probabilities and protein networks with greatly improved confidence compared to previous methods.

Increased coverage of PHFiller using genome context data. PHFiller, a previous algorithm developed by the Karp lab for filling pathway holes, utilized homology and pathway based evidence to determine the probability that a candidate enzyme filled a particular pathway hole (Green et al. 2004). Candidate enzymes for reaction R were identified by searching the organism's genome for homologs of a set of isozyme sequences from other organisms that catalyze reaction R. The algorithm does not identify candidates for pathway holes for which no isozyme sequences are available. Approximately 20% (44 pathway holes) of the remaining pathway holes in the CauloCyc Pathway/Genome Database (PGDB) are reactions for which such sequences are unavailable. We have increased the coverage of the PHFiller algorithm to include these reactions by incorporating genome context data into its search for candidate enzymes.

The protein complex ortholog method, a new source of genome context. In addition to the integration of phylogenetic profiles and gene neighborhood methods into the PHFiller algorithm, the Karp lab has developed an algorithm for identifying functionally associated gene pairs based on known protein complexes. If genes A and B in organism I are known to participate in a protein complex, then we infer that their orthologs, A' and B' in organism II, are functionally related. The EcoCyc PGDB describes 247 heterocomplexes, which yield almost 1400 protein pairs that participate in those complexes. The CauloCyc PGDB includes 158 protein pairs that are orthologs of the *E. coli* proteins pairs. Of these 158 pairs, 122 are annotated with the same COG functional category (Tatusov et al. 2003) and only 60 of these pairs have been identified with high confidence (confidence score greater than 0.7) in the STRING database (von Mering et al. 2003), indicating that our new method finds new functional relationships.

Integrated Protein Interaction Networks for 230 Microbes. The McAdams lab has combined four different types of functional genomic data to create high coverage protein interaction networks for 230 microbes. The integration algorithm naturally handles statistically dependent predictors and automatically corrects for differing noise levels and data corruption in different evidence sources. We find that many of the predictions in each integrated network hinge on moderate but consistent evidence from multiple sources rather than strong evidence from a single source, yielding novel biology which is missed if a single data source such as coexpression or coinheritance is used in isolation. In addition to statistical analysis and recapitulation of known biology, we demonstrate that these subtle interactions can discover new aspects of even well studied functional modules, such as the flagellar hierarchy and the cell division apparatus. This analysis has produced the largest collection of probabilistic protein interaction networks compiled to date, and the methods can be applied to any sequenced organism and any kind of experimental or computational technique which produces pairwise measures of protein interaction.

References

1. Green, M. L. and P. D. Karp (2004). "A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases." *BMC Bioinformatics* 5: 76.
2. Tatusov, R. L., N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, S. Smirnov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin and D. A. Natale (2003). "The COG database: an updated version includes eukaryotes." *BMC Bioinformatics* 4: 41.
3. von Mering, C., M. Huynen, D. Jaeggi, S. Schmidt, P. Bork and B. Snel (2003). "STRING: a database of predicted functional associations between proteins." Nucleic Acids Res 31(1): 258-61.

6

Pipelining RDP Data to the Taxomatic and Linking to External Data Resources

S. H. Harrison[1], P. Saxman[1], T.G. Lilburn[2], J.R. Cole[1], and **G.M. Garrity**[1]* (garrity@msu.edu)

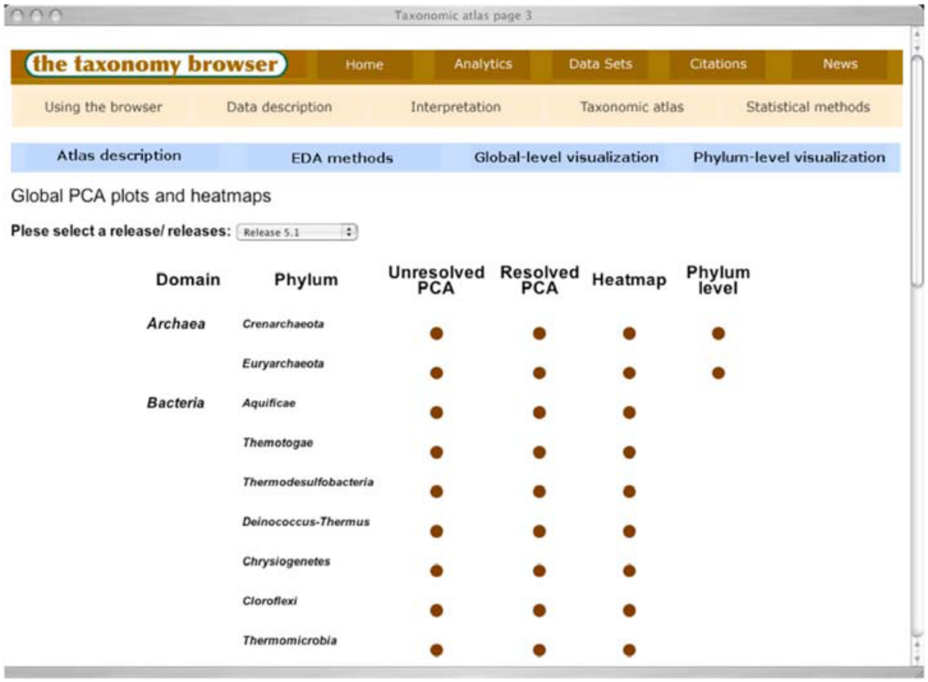
[1]Michigan State University, East Lansing, MI and [2]American Type Culture Collection, Manassas, VA

The taxonomic atlas represents an ongoing experiment in visualization of evolutionary relationships among the prokaryotes. Starting at a point of interest, the system allows users to move through a hierarchical classification, at different levels of taxonomic resolution, so that they may better gauge relationships based on a given gene, group of genes, or other quantitative signal that they might deem relevant. To demonstrate the potential of the methodology, we developed a data-driven atlas of taxonomic/phylogenetic heatmaps, based on a nomenclatural taxonomy that came to be known as the "Taxomatic". Since its inception, the prototype website has been moved to a production web server (http://taxoweb.mmg.msu.edu) where it is maintained and periodically updated. Methods were also developed to permit retrieval and side-by-side viewing of multiple interactive PCA plots from different releases of the underlying taxonomies to permit end-users to readily visualize changes that might otherwise go unrecognized using alternative techniques. For prototyping, we selected Insightful's StatServer as our platform for deploying our interactive graphics to the user community, but have reached the limits of this technology. In order to significantly reduce the amount of time needed to serve up the maps in the atlas, we have turned to AJAX-based technologies to develop a phylogenetic mapping service similar to the well-known geographical mapping service provided by Google. This provides server-side on-the-fly generation of image-mapped files that will ultimately link to key references in the literature, external sets of sequence and phenotypic data, and sources of viable cultures, when available. Together with browser-side JavaScript, these technologies can duplicate most (but not all) of the functionality available in Graphlets, but without the inherent client-side problems.

Using our tools, end-users can browse a taxonomic hierarchy, display, zoom, and re-center the view of a heatmap, highlight and display the names of the higher-level taxa fully visible on the current view, and adjust the taxonomic hierarchy to the show the taxon selected (clicked) on the heatmap by the user. At full zoom, it is possible to adjust the taxonomic hierarchy to identify the organism(s) selected (clicked) on the heatmap and link to external resources.

To build and maintain a carefully annotated and up-to-date reference classification we are connecting the "Taxomatic" to the RDP-II's database and tools, which will ensure timely gathering and rapid alignment of sequences. The models we build based on these alignments and provide via the "Taxomatic" must not only keep pace with the relevant sequence databases, but also must themselves be quickly and objectively built, organizing the data into an optimal or near optimal structure. This is being done using the SOSCC algorithm which was developed to automate this task and will play a central role in ensuring that the internal databases of both the "Taxomatic" and the RDP are relatively free of annotation errors and taxonomic anomalies that arise from the widely used nomenclatural model. The SOSCC code has been significantly revised and now supports both hierarchical and non-hierarchical classification techniques and varying levels of classification stringency. The code has been fully documented and has been packaged as an S-Plus library for distribution from the "Taxomatic" web site.

* Presenting author



Global PCA plots and heatmaps

Plese select a release/ releases: Release 5.1

Domain	Phylum	Unresolved PCA	Resolved PCA	Heatmap	Phylum level
Archaea	*Crenarchaeota*	●	●	●	●
	Euryarchaeota	●	●	●	●
Bacteria	*Aquificae*	●	●	●	
	Themotogae	●	●	●	
	Thermodesulfobacteria	●	●	●	
	Deinococcus-Thermus	●	●	●	
	Chrysiogenetes	●	●	●	
	Cloroflexi	●	●	●	
	Thermomicrobia	●	●	●	

Figure 1. A screenshot of the taxonomic atlas provided by the "Taxomatic".

Significant changes are occurring in the manner, style, and pace of scientific publications. Links to datasets and other online resources are becoming increasingly important to the scientific enterprise, yet traversing those links to relevant information and services while avoiding those that succumbed to "link-rot" (e.g. the notorious Error 404 – Link not found" problem) is an increasingly difficult challenge. While the ramifications of this problem have been discussed by Garrity and Lyons, implementation of a satisfactory solution has remained elusive. As an extension to this project we are investigating the feasibility of using interactive heatmaps and other graphics as navigational devices to link to a collection of persistently maintained "mini-monographs", which in turn provide persistent links to other available data, services, and information resources, specific to a given organism. This is being done through the use of N4L information objects that are uniquely and persistently identified with Digital Object Identifiers. By embedding N4L DOIs into our data structures, this will free us from the necessity of constantly updating the taxonomic information tied to each sequence; yet guarantee that the associated information is up-to-date. As the N4L model supports multiple taxonomic views and concepts, it will also provide a mechanism whereby deviations between different models exist, especially those that are constrained by rules of nomenclature. This is helping to identify areas where nomenclatural anomalies remain because rate of sequencing significantly outstrips the pace of taxonomic revision. This approach will also provide a mechanism whereby persistent links can be established to novel evolutionary lineages of critical importance to DOE missions at the earliest possible point in time, well before such lineages are subject to the formal rules of nomenclature.

References

1. Cole, J. R., B. Chai, T. L. Marsh, R. J. Farris, Q. Wang, S. A. Kulam, S. Chandra, D. M. McGarrell, T. M. Schmidt, G. M. Garrity, and J. M. Tiedje. 2003. "The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy," *Nucleic Acids Res* 31:442-3.
2. Garrity, G. M., and C. Lyons. 2003. "Future-proofing biological nomenclature," *OMICS* 7:31-31.
3. Lilburn, T. G., and G. M. Garrity. 2003. "Exploring prokaryotic taxonomy.,"*Int. J. System. Evol. Micro.* 53:7-13
4. Garrity, G. M., and T. G. Lilburn. 2005. "Self-organizing and self-correcting classifications of biological data," *Bioinformatics* 21:2309-14.

* Presenting author

7

Geobacter Project Subproject I: Genome Sequences of *Geobacteraceae* in Subsurface Environments Undergoing *In Situ* Uranium Bioremediation and on the Surface of Energy-Harvesting Electrodes

Jessica Butler* (jbutler@microbio.umass.edu), Regina O'Neil, Dawn Holmes, Muktak Aklujkar, Ray DiDonato, Shelley Haveman, and **Derek Lovley**

University of Massachusetts, Amherst, MA

The overall goal of the Genomics:GTL *Geobacter* Project is to develop genome-based *in silico* models that can predict the growth and metabolism of *Geobacteraceae* under a variety of environmental conditions. These models are required in order to optimize practical applications of *Geobacteraceae* that are relevant to DOE interests. The goal of Subproject I is to determine the genetic potential of the *Geobacteraceae* present in subsurface environments undergoing *in situ* uranium bioremediation and on the surface of energy-harvesting electrodes. This not only provides information on what metabolic modules need to be included in the *in silico* models but makes it possible to monitor the metabolic state and rates of metabolism in diverse environments by measuring transcript levels of key diagnostic genes. In order to be as comprehensive as possible, the plan is to obtain genome sequences from: 1) pure culture isolates for which there is substantial physiological information; 2) isolates from environments of interest that have the 16S rRNA gene sequences identical to those of the *Geobacter* species that predominate in these environments; 3) single cells from the environments of interest; and 4) genomic DNA extracted from the environments of interest. Progress has been made with all four approaches.

The availability of several new *Geobacteraceae* genome sequences has made it possible to further evaluate the diversity within sequences of closely related members of this family. For example, an in-depth comparison of the relatively well-studied genome of *Geobacter sulfurreducens* and the recently completed genome of *Geobacter metallireducens* revealed that only 60% (2131) of all genes were orthologous between the two genomes. The largest region of the *G. metallireducens* genome that is not present in *G. sulfurreducens* encoded genes for the degradation of aromatic compounds, consistent with the capacity for aromatics metabolism by *G. metallireducens*, but not *G. sulfurreducens*. Although both organisms contain genes for ca. 125 *c*-type cytochromes only 55% of the cytochromes in *G. sulfurreducens* have orthologs in the *G. metallireducens* genome. There are many instances of species-specific duplications of cytochrome genes, as well as instances of gain or loss of heme-binding motifs in orthologous cytochromes. Surprisingly, *c*-type cytochromes that have been shown to be important in Fe(III) reduction in *G. sulfurreducens*, such as the outer-membrane cytochromes OmcB, OmcF, OmcG, and OmcS, do not have orthologs in the *G. metallireducens* genome. Periplasmic cytochromes appear to be better conserved. For example, there are orthologs to the two periplasmic cytochromes, PpcA and MacA, previously shown to be important in Fe(III) reduction in *G. sulfurreducens*. There was much higher conservation of cytoplasmic proteins. For example, 87% of the 589 proteins in the current *in silico* model of central metabolism of *G. sulfurreducens* have orthologs in *G. metallireducens*. Of the 136 cytoplasmic genes that appear to be involved specifically in the oxidation of acetate and electron transport to the inner membrane in *G. sulfurreducens*, 92% have orthologs in *G. metallireducens*. This is significant because acetate is the key electron donor driving *in situ* uranium bioremediation and production of electricity. These findings suggest that modeling of the central metabolism of the environmentally relevant *Geobacter* species may benefit from the existing *G. sulfurreducens in silico* model. However, the results also suggest that there may be little conservation among *Geobacter*

* Presenting author

species in the proteins involved in electron transfer through the periplasm and outer membrane, with the exception of the electrically conductive pilin nanowires that are thought to be the conduit for electron flow from the surface of the cell onto Fe(III) oxides. This conclusion was also supported by analysis of the recently completed genome of *Pelobacter carbinolicus* and the draft genome of *Pelobacter propionicus*.

The first milestone towards the goal of sequencing the genomes of predominant *Geobacteraceae* in the environments of interest has been reached. A 4.8 Mbp draft genome sequence is now available for *Geobacter uraniumreducens*, which was isolated from the in situ uranium bioremediation study site in Rifle, Colorado with an environment-simulating medium in which clay-size minerals served as the source of Fe(III). The 16S rRNA gene sequence of *G. uraniumreducens* matches a sequence that predominates during in situ uranium bioremediation, and it is expected that the short interval between isolation and genome sequencing minimized any potential inactivation, loss, and rearrangement of genes that can accompany propagation of a strain in the laboratory. Although *G. uraniumreducens* has genes for 100 c-type cytochromes, only 20 multi-heme cytochromes have orthologs in the *G. sulfurreducens* and *G. metallireducens* genomes. This contrasts with the presence of a high proportion of orthologs for central metabolism and inner membrane electron transport enzymes.

This year, techniques were developed for immediately freezing samples in the field such that the cells remain intact for subsequent cultivation or sequencing of single cells. With this method it has been possible to recover additional organisms from the uranium bioremediation study site with 16S rRNA gene sequences that are identical to those that predominate during in situ uranium bioremediation. Furthermore, using a multiple displacement amplification approach, genomic DNA was amplified from single cells of *Geobacter* species obtained from these samples. We are awaiting the sequencing results.

Quantification of the transcript levels of a broad range of *Geobacteraceae* genes in a diversity of subsurface environments requires information on the sequence heterogeneity among genes that are diagnostic of important metabolic states. Although some information can be obtained from the genome sequencing described above, these approaches can not yet reasonably be applied to a large number of environments or at a large number of time intervals during the in situ uranium bioremediation process. The sequence diversity of key diagnostic genes was characterized in detail for four sedimentary environments. Degenerate PCR primers were designed from the sequences of genes that are highly conserved throughout the range of pure culture *Geobacteraceae*, as well as the genomic DNA from environments in which *Geobacteraceae* predominate. The diversity of 16S rRNA gene sequences detected was extremely low both within and among these sites. *Geobacteraceae* with 16S rRNA gene sequences closely related to the subsurface isolate *Geobacter bemidjiensis* accounted for 50-98% of the microbial community, and these sequences were 97-100% similar to each other. However, other genes amplified with this method had sequence similarities as low as 50-75%. These findings have a significant impact on strategies for evaluating gene expression in *Geobacteraceae*-dominated environments.

8

Geobacter Project Subproject III: Functional Analysis of Genes of Unknown Function

Maddalena Coppi* (mcoppi@microbio.umass.edu), Carla Risso, Gemma Reguera, Ching Leang, Helen Vrionis, Richard Glaven, Muktak Aklujkar, Xinlei Qian, Tunde Mester, and **Derek Lovley**

University of Massachusetts, Amherst, MA

The development of models that can predict the physiological responses of *Geobacteraceae* under different environmental conditions in contaminated subsurface environments or on the surface of energy-harvesting electrodes requires an understanding of the function of the genes expressed in these environments. However, no function has been assigned to a substantial number of genes in *Geobacteraceae* genomes and the actual physiological role of many genes annotated as having specific physiological functions has yet to be assessed.

For example, it is important to understand acetate metabolism in *Geobacter* species, because acetate is the electron donor driving *in situ* uranium bioremediation and electricity production. Analysis of the *Geobacter sulfurreducens* genome revealed the presence of three homologs of the monocarboxylate transporter of *E. coli*, YjcG, which has recently been demonstrated to catalyze sodium-dependent acetate uptake. These homologs are 54-56% similar to *E. coli* YjcG and are *ca.* 90% similar to each other. *G. sulfurreducens* retained the ability to grow on acetate if the transporter genes were deleted singly, but a triple mutant could not be isolated. This result suggests that the three transporters are essential for growth on acetate, but that they may be functionally redundant. We are currently investigating the possibility of compensatory interactions between the three transporters in the mutant strains. The central metabolic pathways for acetate oxidation and incorporation into biomass have also been subjected to intensive genetic analysis. A model of acetate metabolism based on the results of these studies will be presented.

Previous studies demonstrated that at subatmospheric oxygen tensions, *G. sulfurreducens* can grow utilizing oxygen as an electron acceptor, a physiological capability that may be important for survival of *Geobacter* species in subsurface environments. Two complexes potentially involved in oxygen respiration are encoded in the genome of genome of *G. sulfurreducens*, a cytochrome *c* oxidase and a cytochrome *d* oxidase. Both enzymes are present in most aerobic bacteria and have a low or high affinity for oxygen, respectively. A *G. sulfurreducens* mutant lacking the cytochrome *c* oxidase could not grow on oxygen, but retained the ability to consume oxygen, presumably via the cytochrome *d* oxidase. Further genetic analysis of this possibility is underway.

The physiologic role of the SfrAB complex of *G. sulfurreducens*, which was previously designated a cytoplasmic Fe(III) reductase, was investigated in detail, because understanding the site of metal reduction is crucial for modeling the energetics of this process. A knockout mutant deficient in SfrAB could not grow with acetate as the electron donor with either fumarate or Fe(III) as the electron acceptor, but readily grew with hydrogen or formate as the electron donor, if acetate was provided as a carbon source. This phenotype suggested that the SfrAB-deficient mutant was specifically impaired in acetate oxidation via the TCA cycle. After several weeks, SfrAB deficient strains developed the ability to grow on fumarate with acetate serving as the electron donor. Membrane and soluble fractions prepared from these acetate-adapted strains, were depleted of NADPH-dependent Fe(III), viologen, and quinone reductase activities relative to those of wild type. It was hypothesized that the lack of SfrAB inhibits growth on acetate by increasing the NADPH:NADP

* Presenting author

ratio and thereby inhibiting the isocitrate dehydrogenase reaction of the TCA cycle. Comparison of global gene expression profiles in the adapted mutant and wild type strains provided evidence of ATP depletion, impaired amino acid biosynthesis, and decreased rates of acetate uptake, all of which were consistent with a suboptimal rate of acetate oxidation via the TCA cycle in the mutant. In addition, a potential NADPH-dependent ferredoxin oxidoreductase was upregulated in the mutant. These results indicate that SfrAB is not an Fe(III) reductase, but rather might serve as a major route for NADPH oxidation in *G. sulfurreducens*. These findings, coupled with the fact that metabolically active *G. sulfurreducens* spheroplasts were incapable of Fe(III) reduction suggest that cytoplasmic Fe(III) is not an important process in *G. sulfurreducens*, consistent with other recent functional studies.

Functional analyses have also provided new insights into the mechanisms by which *G. sulfurreducens* produces electricity. We have recently discovered that *G. sulfurreducens* expresses electrically conductive pili that appear to serve as electrical conduits from the cell to Fe(III) and Mn(IV) oxides. As reported last year, initial genetic studies suggested that that outer-membrane *c*-type cytochrome, OmcS, mediated electrical contact between *G. sulfurreducens* and that pili were not required for electricity production. These results may have been due to the fact that, in these studies, power production was limited by the rate of electron transfer from the cathode to oxygen in the overlying water. When cathode limitation was eliminated via the use of a potentiostat, current production by wild type *G. sulfurreducens* increased more than 10-fold and the pilus-deficient mutant was found to be significantly impaired, with a power output that was only 17% of that of wild type. Confocal laser scanning microscopy revealed that, in the absence of cathode limitation, the wild-type cells produced thicker biofilms on the anode, whereas the pilus-deficient mutant produced thinner biofilms more similar to those observed in the previously used cathode-limited systems. These results suggest that cells that are not in close contact with the electrode may transfer electrons through the biofilm via the electrically conductive pili. Deletion of a gene designated *gumC* eliminated the production of exopolysaccharide and also negatively impacted power production, reducing it to less then half of wild type levels. These results suggest that exopolysaccharide production also plays an important role in electron transfer to electrodes. A combination of biochemical, genetic and microscopic analyses demostrated that the *gumC* mutant overproduces pili, which may enable it to compensate for the absence of exopolysaccharide production and transfer electrons to insoluble electron acceptors. Production of current by a *gumCpilA* double mutant was less than 10% of wild type.

Numerous other functional genomics studies are underway. For example, genetic analysis of putative metal resistance genes lead to the identification of genes involved zinc, cadmium, and copper resistance. Surprisingly, deletion of one of the copper-resistance genes, *cusA*, also prevented growth with Fe(III) serving as the electron acceptor. Additional outer membrane *c*-type cytochromes have been implicated in electron transfer to Fe(III) citrate, Fe(III) oxides and electrodes by genetic studies, but several of these cytochromes appear to have functions that are not directly related to electron transfer to metals or electrodes. Additional targets for functional analysis have been selected based on higher levels of expression during growth on Fe(III) oxide or electrodes. The function of proteins that form complexes with cytochromes previously shown to be involved in extracellular electron transfer are also being investigated as are those of proteins which are conserved throughout the *Geobacteraceae*, but not found in the genomes of other organisms.

9

Comparative Genomic Analysis of Five *Rhodopseudomonas palustris* Strains: Insights into Genetic and Functional Diversity within a Metabolically Versatile Species

Yasuhiro Oda[1]* (yasuhiro@u.washington.edu), Frank W. Larimer[2], Patrick Chain[3], Stephanie Malfatti[3], Maria V. Shin[3], Lisa M. Vergez[3], Loren Hauser[2], Miriam L. Land[2], Dale A. Pelletier[2], and **Caroline S. Harwood**[1]

[1]University of Washington, Seattle, WA; [2]Oak Ridge National Laboratory, Oak Ridge, TN; and [3]Lawrence Livermore National Laboratory, Livermore, CA

Rhodopseudomonas palustris is a facultatively photosynthetic bacterial species that has the potential to be used as a biocatalyst for hydrogen production, carbon sequestration, biomass turnover, and biopolymer synthesis. The genome of *R. palustris* strain CGA009 has been reported and consists of a 5.46 Mb chromosome with 4836 predicted protein-coding genes. Several studies have shown that the *R. palustris* species is comprised of genetically and phenotypically diverse strains. To identify the core characteristics of the species that are essential for proper physiological functioning and to identify new metabolic capabilities, the DOE Joint Genome Institute sequenced four additional strains of *R. palustris*. These strains, BisB5, HaA2, BisB18, and BisA53, were directly isolated from agar plates that had been inoculated with freshwater sediments. Their 16S rRNA gene sequences differ from that of strain CGA009 by about 2% and their BOX-PCR genomic DNA fingerprint patterns differ significantly. The genome of strain BisB5 consists of a 4.89 Mb chromosome with 4386 predicted genes. Strain HaA2 has a 5.33 Mb chromosome with 4687 predicted genes, strain BisB18 has a 5.51 Mb chromosome with 4949 predicted genes, and strain BisA53 has a 5.50 Mb chromosome with 4913 predicted genes. Approximately 60 to 80% (depending on the strain) of the genes from strain CGA009 were present in each strain, and these may represent the core genes of the *R. palustris* species. However, whole genome comparisons among strains showed a high degree of genome rearrangement in terms of gene orders and reading directions. Furthermore, there were high numbers of genes (250 to 560 genes) that were specific to a given strain and not seen in any other strain. Based on their gene inventories, each strain is predicted to have strain-specific physiological traits. Strain CGA009 is well equipped for nitrogen fixation with three nitrogenase isozymes and four sets of glutamine synthetases, strain BisB5 has expanded anaerobic aromatic degradation capabilities (e.g., phenylacetate degradation), strain HaA2 should be well-adapted for growth in oxygen as it encodes seven different aerobic terminal oxidases, and strain BisB18 should be able to grow well anaerobically in dark (e.g., carbon-monoxide dehydrogenase genes, three sets of pyruvate-formate lyase genes, formate-hydrogen lyase genes, and DMSO reductase genes). Finally, strain BisB53 has an expanded set of exopolysaccharide synthesis genes and readily attaches to surfaces to form biofilms. Despite these differences, there were relatively few obvious examples of lateral gene transfer in the genomes and the genomes harbor relatively few insertion sequences or transposons compared to other bacterial species. Our comparative genomic analysis suggests that *R. palustris* is a dynamic species comprised of diverse ecotypes that are well adapted to specific environmental niches.

* Presenting author

10

Functional Analysis of *Shewanella*, a Cross Genome Comparison

Margrethe H. Serres* (mserres@mbl.edu) and **Monica Riley** (mriley@mbl.edu)

Marine Biological Laboratory, Woods Hole, MA

Shewanella oneidensis MR-1 was initially chosen as a model organism by the Department of Energy (DOE) based on its unique metal reducing and bioremediation capabilities. Data generated from the analyses of various *Shewanella* strains showed that this genus contained species adapted to life in a variety of environmental niches including land, lake sediments, fresh water and marine environments. The range of environments where *Shewanella* successfully lives implies that it is highly versatile in its capacity to respire, metabolize nutrients, and sense its surroundings for available nutrient sources and electron acceptors. DOE subsequently funded the sequencing of several additional *Shewanella* strains of varying ecological properties, and high quality draft sequences of these genomes have been made available. In our work we analyze the predicted protein sequences from *S. oneidensis* MR-1, 10 strains sequenced at JGI (*S. putrefaciens* CN-32, *S. loihica* (formerly alga) PV-4, *S. baltica* OS155, *S. frigidimarina* NCIMB400, *S. denitrificans* OS217, *S. amazonensis* SB2B, *Shewanella* sp. MR-7, *Shewanella* sp. ANA-3, *Shewanella* sp. MR-4, *Shewanella* sp. W3-18-1), and two environmental samples from the Sargasso Sea (*Shewanella* SAR-1, *Shewanella* SAR2) sequenced by TIGR.

Our research focuses on studying groups of sequence related proteins and whether such groups can give us insight into how organisms adapt to their environments. The duplication of genes followed by diversification of their sequences results in a group of proteins encoding similar or related functions. This process is believed to be an important means of functional specialization and adaptation. The *Shewanella* genome sequences provide an excellent resource to study the distribution of protein groups and to analyze the activities they encode in order to find evidence of specialization of functions related to their metabolic capabilities and their environmental phenotypes.

Pair-wise alignments of the protein sequences encoded by the *Shewanella* genomes were produced using the AllAllDb program of Darwin (Data Analysis and Retrieval With Indexed Nucleotide/peptide sequence package), version 2.0. Fused proteins (arising from gene fusion events) were separated into smaller proteins corresponding to their un-fused components. Protein groups were then generated from the pair-wise alignments in a transitive grouping process. Two methods were used to compare protein groups across the 13 *Shewanella* genomes. In one method groups were initially generated from the proteins encoded by *S. oneidensis* MR-1. , and these groups were further used to search for sequence similar matches in the other 12 genomes. Sequence similarities of 175 PAM units or less and alignments over at least 45% of the protein sequences were applied for this comparison. A total of 406 protein groups containing 2 or more *S. oneidensis* MR-1 proteins were compared this way. In the second method sequence related groups were generated directly from the pair-wise alignments of the proteins from the 13 *Shewanella* strains. The same transitive grouping process was applied, but the sequence similarities were restricted to 125 PAM units or less over 70% of the protein sequences. In the second method 4702 protein groups were generated.

We are analyzing the protein groups for their distribution among the 13 *Shewanella* strains and for the functions they encode. Groups of proteins with functions relating to anaerobic respiration, chemotaxis and environmental sensing will be presented.

* Presenting author

11

Comparative Genomic and Proteomic Insight into the Evolution and Ecophysiological Speciation in the *Shewanella* Genus

James M. Tiedje*[1] (tiedjej@msu.edu), Konstantinos T. Kostantinidis[1], Joel A. Klappenbach[1], Jorge L.M. Rodrigues[1], Mary S. Lipton[4], Margaret F. Romine[4], Sean Conlan[10], LeeAnn McCue[4], Patrick Chain[6], Anna Obraztsova[3], Loren Hauser[2], Margrethe Serres[7], Monica Riley[7], Carol S. Giometti[5], Eugene Kolker[8], Jizhong Zhou[2], Kenneth H. Nealson[3], and James K. Fredrickson[4]

[1]Michigan State University, East Lansing, MI; [2]Oak Ridge National Laboratory, Oak Ridge, TN; [3]University of Southern California, Los Angeles, CA; [4]Pacific Northwest National Laboratory, Richland, WA; [5]Argonne National Laboratory, Argonne, IL; [6]Lawrence Livermore National Laboratory, Livermore, CA; [7]Marine Biological Laboratory, Woods Hole, MA; [8]BIATECH Institute, Bothell, WA; [9]University of Oklahoma, Norman, OK; and [10]Wadsworth Center, Troy, NY

Members of the genus *Shewanella* are found in a variety of environments, such as freshwater lakes, marine sediments, subsurface formations, and at variable depths in redox stratified aquatic systems. The ecological and physiological diversities among species of this genus suggest a high degree of specialization. The mechanisms and factors that drive speciation are still not well understood. For this reason, we are taking advantage of full genome sequencing of 17 *Shewanella* species and aim to: 1) define the genetic differences and mechanisms that account for the ecological success in different environments and support different physiologies, 2) identify mechanisms of evolution and speciation for *Shewanella* species, and 3) determine a set of core genes important for metal reduction. Results from four genomic sequences (*S. frigidimarina*, *S. putrefaciens* CN-32, S. sp. PV-4, and *S. denitrificans*) compared to the finished genome of *S. oneidensis* MR-1 indicated the presence of 10,000 unique genes, while only 2,200 genes are shared among all sequenced species. Each sequenced species contained a subset of genome specific genes (25%), with a substantial number of those (16%) annotated as hypothetical open reading frames (ORF). The pair-wise genetic distances using the average nucleotide identity (ANI) of all genes according to the method of Kostantinidis and Tiedje (2005) between these 4 genomes were approximately 70%, indicating that these genomes are not closely related, and hence a species might harbor several distinct ecotypes, appropriately evolved for a specific environment. The above 4 genomes together with an additional 7 that have been more recently sequenced to offer resolution within species present an unprecedented evolutionary gradient for study of the diversification of a bacterial genus at varied level of resolution. The sequenced genomes with various degree of relatedness within a single genus will provide ideal model system for studying evolutionary processes and forces such as positive, neutral and negative selection in prokaryotes (Figure 1).

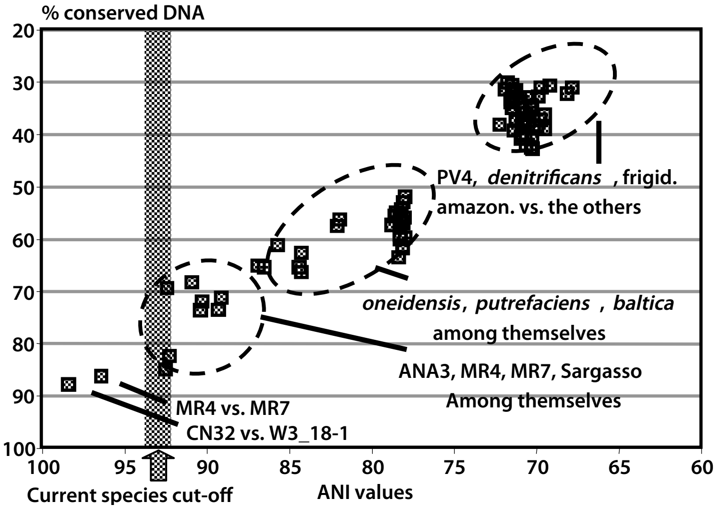


Fig. 1. Conserved gene core vs. evolutionary distance.

* Presenting author

Insights from the analysis of 11 *Shewanella* genomes indicated that 2,000-2,200 genes are shared with *Vibrio*, a number higher than those shared within the *Shewanella* genus. This finding demonstrates the challenges of identifying ecologically relevant genes, based solely on sequence analysis. Thus, we are analyzing the proteomes of all different species to identify a core set of expressed proteins under defined conditions. Whole cell lysates from 13 *Shewanella* species were analyzed by two-dimensional electrophoresis (2DE) and the patterns were compared. Sorting the 2DE patterns by constellations of similar spots resulted in grouping of the species in close approximation to the *gyrB*-based phylogenetic tree. We are studying the distribution of sequence similar protein families in the 13 *Shewanella* genomes to detect functional adaptation relating to their environments. We are also examining the species distribution of cytochromes, short repeats, and IS elements, as well as conservation of regulatory elements for insights into the evolution of these species. In addition, physiological studies for different strains are underway, testing salinity, temperature, pH, carbon sources under aerobic conditions, a variety of electron acceptors with lactate or N-acetyl glucosamine as electron donor, ability of producing current in a mediator-less biofuel cell, and growth on Cr(VI). Physiological studies will be correlated to the genomic and proteomic content and variability in all species.

Reference

1. Kostantinidis, K. and J.M. Tiedje. 2005. "Genomic insights that advance the species definition for prokaryotes." *Proc. Natl. Acad. Sci. U.S.A.* 7:2567-2572.

12

Microbial Genome Sequencing with Ploning from the Wild: A Progress Report

Kun Zhang[1]* (kzhang@genetics.med.harvard.edu), Adam C. Martiny[2], Nikkos B. Reppas[1], Kerrie W. Barry[3], Joel Malek[4], Sallie W. Chisholm[2], and **George M. Church**[1]

[1]Harvard Medical School, Boston, MA; [2]Massachusetts Institute of Technology, Cambridge, MA; [3]Joint Genome Institute, Walnut Creek, CA; and [4]Agencourt Bioscience, Beverley, MA

With less than 1% of microorganisms easily cultured, obtaining genome sequence and continuity for the remaining 99% has been one of the greatest challenges in environmental genomic studies. We aimed at developing a method, polymerase cloning (ploning) for genome sequencing directly from single uncultured cells, and applying the method to study the Prochlorococcus community structure in open oceans.

The first critical component of ploning is to perform whole genome amplification on a single template molecule. We have reported a real-time isothermal amplification system that successfully addresses the issue of background non-specific amplification in last year's meeting. We also investigated amplification bias in two *E. coli* polymerase clones (plones) using Affymetrix chip hybridization, and showed that bias is randomly distributed.

Here we present several recent progresses in the second critical component of our method: genome sequencing from plones. We have prepared several plones from single Prochlorococcus cells of the MIT 9312 strain. Our initial attempt of performing shotgun sequencing on such plones failed because of issues in sequencing library construction, such as abnormal insert size, high vector content and low cloning efficiency. We hypothesized that such problems are due to the high order DNA

* Presenting author

branching structure generated by multiple displacement amplification. A three-step enzymatic treatment has been developed to resolve the high order DNA structure, so that the chimeric rate has been reduced from as high as 52% to 6%. We have sequenced two MIT 9312 plones, one at the sequencing depth of 3.5x and the other at 4.7x, and recovered 62% and 66% of the genome respectively. Full genome coverage can be achieved by increasing the sequencing depth to ~24x or PCR sequencing from the plones. The estimated mutation rate in single cell amplification is $<2\times10^{-5}$.

When applying the ploning method to ocean samples collected from Hawaii, we encounter another technical problem: the amount of cell-free DNA is more than that within a live cell in single-cell dilution. We have developed a DNase-based protocol to remove contamination of cell-free DNA. We are screening for good *Prochlorococcus* plones from ocean samples, and will sequence a few in the near future.

13

Application of a Novel Genomics Technology Platform

Karsten Zengler[1], Marion Walcher[1], Carl Abulencia[1], Denise Wyborski[1], Trevin Holland[1], Fred Brockman[2], Cheryl Kuske[3], and **Martin Keller**[1]* (mkeller@diversa.com)

[1]Diversa Corporation, San Diego, CA; [2]Pacific Northwest National Laboratory, Richland, WA; and [3]Los Alamos National Laboratory, Los Alamos, NM

A technology platform has been developed to obtain whole genome sequences from targeted uncultured microorganisms. Our approach combines fluorescence in situ hybridization (FISH) followed by amplification of whole genomes using multiple displacement amplification (MDA). Microcolonies in microcapsules derived from high throughput cultivation (HTC) or individual cells from the environment are specifically stained with fluorescently labeled oligonucleotide probes targeting 16S rRNA. Target cells are further isolated from non-target microorganisms by flow cytometry. Genomes of these isolated cells are subsequently amplified by whole genome amplification.

The genome from one *Acidobacteria*-microcolony was amplified by MDA. Furthermore, positive MDA products were obtained from 5, 50 and 100 *Acidobacteria* cells and 100 cells affiliated to candidate division TM7 isolated directly from the environment. A small insert library from a MDA product derived from 50 *Acidobacteria* cells has been constructed and a portion of the library (~1000 clones) has been sequenced. The library appears to predominantly contain sequences from an *Acidobacteria* division member that represents a major uncultured subgroup (group #6). Blast scores to published sequences were generally very low, but several clones were directly affiliated to *Solibacter usitatus*, a group 3 Acidobacterium. A total of 82 contigs were assembled which were derived from 982 sequences. The largest of these contigs consists of over 6 kb and was derived from 27 sequences. Three of those sequences are closely related to *Solibacter usitatus*.

* Presenting author