

Joint Meeting

Genomics:GTL Awardee Workshop V

and

Metabolic Engineering Working Group

Interagency Conference on Metabolic Engineering 2007

and

USDA-DOE Plant Feedstock Genomics for Bioenergy Awardee Workshop 2007

Bethesda, Maryland

February 11–14, 2007

Prepared for the
U.S. Department of Energy
Office of Science
Office of Biological and Environmental Research
Office of Advanced Scientific Computing Research
Germantown, MD 20874-1290

<http://genomicsgtl.energy.gov>

Prepared by
Genome Management Information System
Oak Ridge National Laboratory
Oak Ridge, TN 37830
Managed by UT-Battelle, LLC
For the U.S. Department of Energy
Under contract DE-AC05-00OR22725

Welcome

Welcome to the 2007 Joint Genomics:GTL Awardee Workshop, Metabolic Engineering Interagency Working Group Conference, and the USDA-DOE Plant Feedstock Genomics for Bioenergy Joint Program Meeting. Attendees of this year's joint meeting represent a broad cross section of cutting-edge research disciplines, including microbiology, plant biology, genomics, proteomics, physiology, structural biology, metabolic engineering, ecology, evolutionary biology, bioinformatics, and computational biology. Uniting this diverse community of researchers is the shared goal to understand the complex genetic and metabolic systems supporting the life of microbes, plants, and biological communities. This approach is by definition integrative and multidisciplinary, and we applaud the collaborative spirit that brings you to this meeting and informs your research. This joint meeting provides unique opportunities to enhance collaboration and advance science beyond the boundaries of any of these individual research programs.

In May of 2006, the National Research Council of the National Academies completed an independent review of the Genomics:GTL program. The review committee enthusiastically endorsed the systems biology focus of the program, lauded the general excellence of research conducted by its grantees, and re-emphasized the critical importance of the program's core mission—to advance fundamental research on bioenergy solutions, bioremediation of environmental contaminants, and global carbon cycling and sequestration. The review also recommended that, in place of the previously proposed technology-oriented user facilities, the Genomics:GTL program should form interdisciplinary research centers focused on

fundamental research addressing DOE mission needs, including bioenergy. DOE embraced this recommendation, and review of proposals for GTL Bioenergy Research Centers is under way.

This year's meeting features sessions that will be of broad interest to scientists funded by all three of the participating research programs—the Genomics:GTL program, the Metabolic Engineering Working Group, and the USDA-DOE Plant Feedstock Genomics for Bioenergy Joint Program. The goal of the Metabolic Engineering Working Group is the targeted and purposeful alteration of an organism's metabolic pathways to better understand and use cellular pathways for chemical transformation, energy transduction, and supramolecular assembly. The USDA-DOE Plant Feedstock Genomics for Bioenergy joint program focuses on genomics-based research leading to the improved use of biomass and plant feedstocks for the production of such fuels as ethanol or renewable chemical feedstocks. By bringing together these distinct but complementary research communities, we hope to foster the exchange of ideas, sharing of expertise, and formation of new partnerships among researchers.

We look forward to an exciting and productive meeting and encourage you to participate in discussions and build new relationships. We wish to thank the organizers of this year's meeting, especially members of the research community who have generously agreed to play a larger role in organizing this year's breakout sessions. Finally, we thank you for lending your knowledge, creativity, and vision to the represented research programs and wish you continued success in the coming year.



Jerry Elwood
Acting Associate Director
of Science for Biological and
Environmental Research
Office of Science
U.S. Department of Energy



Michael Strayer
Associate Director of
Science for Advanced
Scientific Computing
Research
Office of Science
U.S. Department of Energy



Fred Heineken
Chair of the Interagency
Metabolic Engineering
Working Group
National Science Foundation



Anna Palmisano
Deputy Administrator of
Competitive Programs
Cooperative State Research,
Education, & Extension
Service
U.S. Department of
Agriculture

Agenda

Sunday Evening, February 11th

5:00-8:00 Registration and Poster Setup

6:00-8:00 Mixer

Monday, February 12th

7:00-8:00 Continental Breakfast

8:00-8:45 Welcome

State of Genomics:GTL Program, Vision for the Future
Mike Viola, David Thomassen, and Jim Fredrickson

8:45-9:15 Systems Biology: An Overview

Timothy Donohue – University of Wisconsin, Madison
"Systems Analysis of Solar Energy Stress Responses"

9:15-12:00 Plenary Session

Carbon Processing at Organismal, Community, & Global Scales
Moderator: Dan Drell

Session Objectives: This session will examine the global carbon cycle at various levels of resolution and from a variety of different research perspectives. The goal of the session is to promote discussion of potential contributions of GTL research towards more effectively modeling the global carbon budget and understanding the role of microbes and microbial communities in planetary scale processing and sequestration of carbon.

9:15-9:45 Jae Edmonds – Pacific Northwest National Laboratory
"Biotech, Energy, and a Climate-Constrained World"

9:45-10:15 Bruce Hungate – Northern Arizona University
Title To Be Announced

10:15-10:30 Break

10:30-11:00 Mary Lidstrom – University of Washington
"Integrated Metabolism Studies of Methylophrophy"

11:00-11:30 Ramon Gonzalez – Rice University
"The Role of Metabolic Engineering and Systems Biology in Biofuel Production"

12:00-2:00 Lunch

2:00-4:30 Breakout Session A

USDA-DOE Plant Feedstock Genomics for Bioenergy Joint Program Meeting
Moderators: Ed Kaleikau (USDA) & Sharlene Weatherwax (DOE)

2:00-2:15 Program Overview: Chavonda Jacobs-Young - USDA

2:15-2:30 Vincent Chiang – North Carolina State University

2:30-2:45 Clint Chapple – Purdue University

2:45-3:00 Chang-Jun Liu – Brookhaven National Laboratory

3:00-3:15 Bikram Gill – Kansas State University

3:15-3:30 Richard Dixon – The Nobel Foundation

3:30-4:00 Break

4:00-4:15 William Rooney – Texas A&M University

4:15-4:30 Charles Brummer – University of Georgia

4:30-4:45 John Ralph - USDA-ARS, University of Wisconsin

4:45-5:00 Chris Somerville – Carnegie Institute of Washington

2:00-5:00 Breakout Session B

Data Management & Sharing Tutorial

Organizers: Ed Uberbacher – Oak Ridge National Laboratory
Jeff Grethe – University of California, San Diego

Moderator: John Houghton

2:00-3:30 Panel Discussion (Participants To Be Announced)

3:30-4:00 Break

4:00-5:00 Group Discussion

2:00-5:00 Breakout Session C

Cultivation Methodology Tutorial

Moderator: Joe Graber

Organizers: Slava Epstein & Hans Scholten

2:00-2:20 Slava Epstein – Northeastern University

2:20-2:40 Hans Scholten – Pacific Northwest National Laboratory

2:40-2:55 Martin Keller – Oak Ridge National Laboratory

2:55-3:10 Chris Belnap – University of California, Berkeley

3:10-3:25 Judy Wall – University of Missouri, Columbia

3:30-4:00 Break

4:00-5:00 Group Discussion

5:00-8:00 Poster Session

Tuesday, February 13th

7:00-8:00 Continental Breakfast

8:00-10:00 Plenary Session

Environmental Sensing & Response In Contaminated Habitats
Moderator: Joe Graber

Session Objectives: This session will examine the means by which microorganisms sense environmental conditions and process the acquired information to form an adaptive response. Emphasis will be placed on the role of GTL research in understanding organismal and community dynamics in relevant habitats and how this information can be integrated into the study of contaminant fate and transport.

- 8:00-8:30 Derek Lovley – University of Massachusetts, Amherst
“Adaptive Evolution and Adaptive Responses in Gene Expression of *Geobacter* Species During Uranium Bioremediation and in Microbial Fuel Cells”
- 8:30-9:00 Tim Gardner, Boston University
“Genome-scale mapping of transcription networks in *Shewanella oneidensis*: identification of carbon and metal-respiratory regulation”
- 9:00-9:30 Jill Banfield - University of California, Berkeley
Title To Be Announced
- 9:30-10:00 Adam Arkin – Lawrence Berkely National Laboratory
Title To Be Announced

10:00-10:15 Break

10:15-10:45 Information Representation, Integration, & Sharing Working Group Report

Genomics:GTL Data & Information
Sharing Policy
Jim Fredrickson & Roland Hirsch

10:45-11:15 Report of the Joint Genome Institute

Edward Rubin - Joint Genome Institute

11:15-12:15 Keynote Lecture on Systems Biology

Marc Vidal – Harvard University
“Interactome Networks”

12:15-2:00 Lunch

2:00-5:00 Breakout Session D

Metabolic Engineering Interagency Working Group Conference:
Methods to Examine Biochemical & Metabolic Networks
Organizers: Gail McLean (USDA) & Sharlene Weatherwax (DOE)
Moderators: Mark Segal (EPA) & Fred Heineken (NSF)

- 2:00-3:30 Presentations:
Bernhard Palsson – University of California, San Diego
George Bennett – Rice University
James Liao – University of California, Los Angeles
Jackie Shanks – Iowa State University
- 3:30-4:00 Break
- 4:00-5:00 Group Discussion

2:00-5:00 Breakout Session E

Tools for the Prediction & Validation of Gene Function
Organizers: Margaret Romine – Pacific Northwest National Laboratory
Andrei Osterman – The Burnham Institute
Moderator: Marvin Stodolsky

- 2:00-3:30 Panel Discussion:
Scott Lesley - Scripps Research Institute
Tom Squier - Pacific Northwest National Laboratory
Tim Gardner - Boston University
- 3:30-4:00 Break
- 4:00-5:00 Group Discussion

2:00-5:00 Breakout Session F

Imaging Techniques Tutorial
Organizers: Eva Nogales - University of California, Berkeley
Gang Bao – Georgia Institute of Technology
Moderator: Arthur Katz

- 2:00-2:20 Eva Nogales - Lawrence Berkeley National Laboratory and University of California, Berkeley
- 2:20-2:40 Grant Jensen - California Institute of Technology
- 2:40-2:55 David Spector - Cold Spring Harbor Laboratory
- 2:55-3:10 Hedi Mattoussi - Naval Research Laboratory
- 3:10-3:25 Guido Gaietta - University of California, San Diego
- 3:30-4:00 Break
- 4:00-5:00 Group Discussion

5:00-8:00 Poster Session

Wednesday, February 14th

7:00-8:00 Continental Breakfast

8:00-10:45 Plenary Session

Protein Structure & Function

Moderator: Roland Hirsch

Session Objective: This session will discuss the critical role of protein structural determination in understanding the functional capabilities of gene products. The goal of this session is to promote discussion of new advances in structural biology of proteins and ways in which the GTL community can incorporate these approaches into new research directions.

8:00-8:30 David Eisenberg – University of California, Los Angeles
“Structural Biology: An Overview of Activities at the UCLA-DOE Institute for Genomics & Proteomics”

8:30-9:00 Sunney Xie – Harvard University
“Single-Molecule Approach to Bacterial Molecular Biology: From In Vitro to In Vivo Studies”

9:00-9:30 J.H. David Wu – University of Rochester
Structural Biology of the Cellulosome

9:30-10:00 Keith Hodgson - Stanford Linear Accelerator Center
“Studying Biocomplexity at Nano to Subnanometer Resolution: The Role of Synchrotron Radiation and the DOE National User Facilities”

10:00-10:15 Break

10:15-10:45 Himadri Pakrasi - Washington University
“Membrane Biology Grand Challenge: Engagement of a DOE National User Facility for Integrative Studies of an Energy Producing Photosynthetic Bacterium”

**10:45-11:15 Reports on Tutorial Sessions
(10 minutes each)**

11:15-11:30 Closing Remarks

Contents

Welcome to GTL-MEWG Workshop	iii
---	------------

Agenda.....	v
--------------------	----------

Workshop Abstracts	1
---------------------------------	----------

GTL Milestone 1	3
------------------------------	----------

Section 1: Organism Sequencing, Annotation, and Comparative Genomics.....	3
--	----------

Poster

1 Genomic Reconstruction and Experimental Validation of Catabolic Pathways in <i>Shewanella</i> Species.....	3
Andrei Osterman* (osterman@burnham.org), Dmitry Rodionov, Chen Yang, Yanbing Wang, Margaret Romine, Anna Obratzova,* and Kenneth Nealson	
2 Evolutionary Analysis of Proteins Deduced from 10 Fully Sequenced <i>Shewanella</i> Genomes.....	4
N. Maltsev (maltsev@mcs.anl.gov), D. Sulakhe, A. Rodriguez, M. Syed, and M. Romine	
3 Modeling Conserved Indels as Phylogenetic Markers in <i>Shewanella</i>	7
John P. McCrow* (mccrow@usc.edu), Kenneth H. Nealson, and Michael S. Waterman	
4 <i>Shewanella</i> Population Comparative Genomics and Proteomics: Connecting Speciation, Ecophysiology, and Evolution	8
Jorge L.M. Rodrigues* (rodrig76@msu.edu), Konstantinos T. Kostantinidis, Margaret F. Romine, Margrethe H. Serres, Lee Ann McCue, Mary S. Lipton, Carol S. Giometti, Anna Obratzova, Matt Marshall, Miriam Land, Kenneth H. Nealson, James K. Fredrickson, and James M. Tiedje	
5 The Complete Genome of the Uncultivated Ultra-Deep Subsurface Bacterium <i>Desulforudis audaxviator</i> Obtained by Environmental Genomics	10
Dylan Chivian* (DCChivian@lbl.gov), Eric J. Alm, Eoin L. Brodie, David E. Culley, Thomas M. Gihring, Alla Lapidus, Li-Hung Lin, Steve Lowry, Duane P. Moser, Paul Richardson, Gordon Southam, Greg Wanger, Lisa M. Pratt, Adam P. Arkin (aparkin@lbl.gov), Terry C. Hazen, Fred J. Brockman, and Tullis C. Onstott	
6 Genomic Comparisons Between a Metal-Resistant Strain of <i>Desulfovibrio vulgaris</i> and the Type Strain <i>D. vulgaris</i> Hildenborough	11
C.B. Walker, D. Joyner, D. Chivian, S.S. Stolyar, K. Hillesland, J. Gabster, P. Dehal, M. Price, T.C. Hazen, A.P. Arkin (aparkin@lbl.gov), P.M. Richardson, D. Bruce, and D.A. Stahl*	

<u>Poster</u>	<u>Page</u>
7 Web Tools for Revealing Relationships Among Strains, Taxa, and Communities	12
T.G. Lilburn, S.H. Harrison,* J.R. Cole, P.R. Saxman, and G.M. Garrity (garrity@msu.edu)	
8 High Quality Microbial Finishing at JGI	14
Alla Lapidus* (alapidus@lbl.gov), Eugene Goltsman, Steve Lowry, Hui Sun, Alicia Clum, Stephan Trong, Pat Kale, Alex Copeland, Patrick Chain, Cliff Han, Tom Brettin, Jeremy Schmutz, and Paul Richardson	
9 Evolution of Energy Metabolism in the <i>Geobacteraceae</i>	15
J.E. Butler* (jbutler@microbio.umass.edu), N.D. Young, D. Kulp, and D.R. Lovley	
10 Establishing Potential Chloroplast Function Through Phylogenomics	16
Sabeeha Merchant* (merchant@chem.ucla.edu), Steven Karpowicz, Arthur Grossman, Simon Prochnik, and Dan Rokhsar	
11 Beneficial Effects of Endophytic Bacteria on Biomass Production by Poplar	17
Safiyh Taghavi and Daniel van der Lelie* (vdlelie@bnl.gov)	
Section 2: Microbial Community Sequencing and Analysis	20
12 Structure and Dynamics of Natural Low-Diversity Microbial Communities	20
Jillian F. Banfield* (jill@eps.berkeley.edu), Vincent Denef, Nathan VerBerkmoes, Paul Wilmes, Gene Tyson, John Eppley, Genevieve DiBartolo, Daniela Goltsman, Anders Andersson, Chris Belnap, Brett J. Baker, Linda Kalnejais, A. Pepper Yelton, D. Kirk Nordstrom, Eric E. Allen, Rachel Whitaker, Sheri Simmons, Manesh Shah, Michael Thelen, Gary Andersen, and Robert Hettich	
13 A Novel Binning Approach and Its Application to a Metagenome From a Multiple Extreme Environment	22
N. Maltsev* (maltsev@mcs.anl.gov), M. Syed, A. Rodriguez, B. Gopalan, and F. Brockman	
14 Insights into Stress Ecology and Evolution of Microbial Communities from Uranium-Contaminated Groundwater Revealed by Metagenomics Analyses	25
Christopher L. Hemme,* Ye Deng, Terry Gentry, Liyou Wu, Matthew W. Fields, David Bruce, Chris Detter, Kerrie Barry, David Watson, Paul Richardson, James Bristow, Terry C. Hazen, James Tiedje, Eddy Rubin, Adam P. Arkin (aparkin@lbl.gov), and Jizhong Zhou	
15 Changes in Microbial Community Structure During Biostimulation for Uranium Reduction at Different Levels of Resolution	26
C. Hwang,* W.-M.Wu, T.J. Gentry, J. Carley, S.L. Carroll, D. Watson, P.M. Jardine, J. Zhou, T.C. Hazen, E.L. Brodie, Y.M. Piceno, G.L. Andersen, E.X. Perez, A. Masol, C.S. Criddle, and M.W. Fields	
16 VIMSS Applied Environmental Microbiology Core Research on Stress Response Pathways in Metal-Reducers	28
Terry C. Hazen,* Carl Abulencia, Gary Anderson, Sharon Borglin, Eoin Brodie, Steve van Dien, Matthew Fields, Jil Geller, Hoi-Ying Holman, Richard Phan, Eleanor Wozei, Janet Jacobsen, Dominique Joyner, Romy Chakraborty, Martin Keller, Aindrila Mukhopadhyay, David Stahl, Sergey Stolyar, Judy Wall, Huei-che Yen, Grant Zane, Jizhong Zhou, E. Hendrickson, T. Lie, J. Leigh, and Chris Walker	

<u>Poster</u>	<u>Page</u>
17 Microarrays + NanoSIMS: Linking Microbial Identity and Function	31
Jennifer Pett-Ridge* (pettridge2@llnl.gov), Peter K. Weber, Paul Hoeprich, Philip Banda, Ian Hutcheon, Eoin Brodie, and Gary Andersen	
18 NanoSIMS Analyses of Molybdenum Indicate Nitrogenase and N-Fixation Activity in Diazotrophic Cyanobacteria	32
Jennifer Pett-Ridge, Juliette Finzi, Ian D. Hutcheon (hutcheon1@llnl.gov), Doug Capone, and Peter K. Weber*	
19 Application of a Novel Genomics Technology Platform	34
Mircea Podar, Carl Abulencia, Don Hutchinson, Joseph Garcia, Lauren Hauser, Cheryl Kuske, and Martin Keller* (kellerm@ornl.gov)	
20 Genome-Scale Analysis of the Physiological State of <i>Geobacter</i> Species During <i>In Situ</i> Uranium Bioremediation	35
Dawn E. Holmes* (dholmes@microbio.umass.edu), Regina A. O'Neil, Milind A. Chavan, Muktak Aklujkar, and Derek R. Lovley	
Section 3: Protein Production and Characterization	37
21 High Throughput Selection of Affinity Reagents	37
Peter Pavlik, Nileena Velappan, Hugh Fisher, Csaba Kiss, Minghua Dai, Emanuele Pesavento, Leslie Chasteen, and Andrew Bradbury* (amb@lanl.gov)	
22 Progress on Fluorobodies	38
Nileena Velappan, Hugh Fisher, Csaba Kiss, Minghua Dai, Emanuele Pesavento, Leslie Chasteen, Peter Pavlik, and Andrew Bradbury* (amb@lanl.gov)	
23 High Throughput Screening of Affinity Reagents: Eliminating Helper Phage from Phage Display by the Use of Helper Plasmids	39
Leslie Chasteen, Joanne Ayriss, Nileena Velappan, Peter Pavlik, and Andrew Bradbury* (amb@lanl.gov)	
24 Selecting Affinity Reagents which Recognize Specific Post-Translational Modifications Independently of Sequence Context: The Sulfotyrosine Example	40
John Kehoe, Jytte Rasmussen, Monica Walbolt, Jianlong Lou, James D. Marks, Peter Pavlik, Carolyn Bertozzi, and Andrew Bradbury* (amb@lanl.gov)	
25 A Total Chemical Synthesis Approach to Protein Structure and Function	41
Stephen Kent* (skent@uchicago.edu), Duhee Bang, Thomas Durek, Zachary Gates, Erik Johnson, Brad Pentelute, and Vladimir Torbeev	
26 A Combined Informatics and Experimental Strategy for Improving Protein Expression	43
Osnat Herzberg, John Moulton* (moulton@umbi.umd.edu), Fred Schwarz, and Harold Smith	

<u>Poster</u>	<u>Page</u>
27 Structural and Functional Characterization of a Periplasmic Sensor Domain from <i>Geobacter sulfurreducens</i> Chemotaxis Protein: A Novel Structure from a Family of Sensors in <i>Geobacteraceae</i>	44
P. Raj Pokkuluri, Yuri Y. Londer, Norma Duke, Stephan Wood, Miguel Pessanha, Teresa Catarino, Carlos A. Salgueiro, and Marianne Schiffer* (mschiffer@anl.gov)	
28 High-Throughput Production and Analyses of Purified Proteins	46
F. William Studier* (studier@bnl.gov), John C. Sutherland, Lisa M. Miller, Hui Zhong, and Lin Yang	
Section 4: Molecular Interactions	47
29 Molecular Assemblies, Genes, and Genomics Integrated Efficiently: MAGGIE	47
John A. Tainer* (jat@scripps.edu)	
30 The MAGGIE Project: Identification and Characterization of Native Protein Complexes and Modified Proteins from <i>Pyrococcus furiosus</i>	49
Angeli Lal Menon* (almenon@uga.edu), Farris L. Poole II, Aleksandar Cvetkovic, Saratchandra Shanmukh, Joseph Scott, Francis E. Jenney Jr., Sunia Trauger, Ewa Kalisiak, Gary Siuzdak, Greg Hura, John A. Tainer, and Michael W. W. Adams	
31 The MAGGIE Project: Production and Isolation of Tagged Native/Recombinant Multiprotein Complexes and Modified Proteins from Hyperthermophilic <i>Sulfolobus solfataricus</i>	51
Denise Munoz, Jill Fuss, Kenneth Stedman, Michael W. W. Adams, Gary Siuzdak, Nitin S. Baliga, Steven R. Holbrook, John A. Tainer, and Steven M. Yannoni* (SMYannoni@lbl.gov)	
32 Protein Complex Analysis Project (PCAP): Project Overview	52
Dwayne Elias, Swapnil Chhabra, Jil T. Geller, Hoi-Ying Holman, Dominique Joyner, Jay Keasling, Aindrila Mukhopadhyay, Mary Singer, Tamas Torok, Judy Wall, Terry C. Hazen, Gareth Butland, Ming Dong, Steven C. Hall, Bing K. Jap, Jian Jin, Susan J. Fisher, Peter J. Walian, H. Ewa Witkowska, Lee Yang, Mark D. Biggin* (mdbiggin@lbl.gov), Manfred Auer, Agustin Avila-Sakar, Florian Garczarek, Robert M. Glaeser, Jitendra Malik, Eva Nogales, Hildur Palsdottir, Jonathan P. Remis, Dieter Typke, Kenneth H. Downing, ^a Steven S. Andrews, Adam P. Arkin, Steven E. Brenner, Y. Wayne Huang, Janet Jacobsen, Keith Keller, Ralph Santos, Max Shatsky, and John-Marc Chandonia	
33 Protein Complex Analysis Project (PCAP): Multi-Protein Complex Purification and Identification by Mass Spectrometry	53
Gareth Butland, Ming Dong, Steven C. Hall, Bing K. Jap, Jian Jin, Susan J. Fisher, Peter J. Walian, H. Ewa Witkowska, Lee Yang, and Mark D. Biggin* (mdbiggin@lbl.gov)	
34 Protein Complex Analysis Project (PCAP): Imaging Multi-Protein Complexes by Electron Microscopy	56
Manfred Auer, Agustin Avila-Sakar, David Ball, Florian Garczarek, Robert M. Glaeser, Jitendra Malik, Eva Nogales, Hildur Palsdottir, Jonathan Remis, Max Shatsky, Dieter Typke, and Kenneth H. Downing* (KHDowning@lbl.gov)	

<u>Poster</u>	<u>Page</u>
35 Protein Complex Analysis Project (PCAP): Microbiology Subproject	58
Hoi-Ying Holman, Jay Keasling, Aindrila Mukhopadhyay, Swapnil Chhabra, Jil T. Geller, Mary Singer, Dominique Joyner, Tamas Torok, Judy Wall, Dwayne A. Elias, and Terry C. Hazen* (tchazen@lbl.gov)	
36 Protein Complex Analysis Project (PCAP): High Throughput Strategies for Tagged-Strain Generation in <i>Desulfovibrio vulgaris</i>	60
Swapnil Chhabra* (SRChhabra@lbl.gov), Gareth Butland, Dwayne Elias, Aindrila Mukhopadhyay, John-Marc Chandonia, Jay Keasling, and Judy Wall	
37 Protein Interaction Reporters and Outer Membrane Cytochrome C	61
James E. Bruce* (james_bruce@wsu.edu), Haizhen Zhang, Natalia Zakharova, Xiaoting Tang, Gerhard R. Muske, Liang Shi, James K. Fredrickson, Nikola Tolic, and Gordon A. Anderson	
38 Technologies for Comprehensive Protein Production	62
Sarah Giuliani, Elizabeth Landorf, Terese Peppler, and Frank Collart* (fcollart@anl.gov)	
39 Next-Generation Cell-Permeable Multiuse Affinity Probes (MAPs) and Cognate Tags: Applications to Bioenergy and Metabolic Engineering	63
M. Uljana Mayer, Baowei Chen, Haishi Cao, Ting Wang, Ping Yan, Yijia Xiong, Liang Shi, and Thomas C. Squier* (thomas.squier@pnl.gov)	
40 Advanced Data Analysis Pipeline for Determination of Protein Complexes and Interaction Networks at the Genomics:GTL Center for Molecular and Cellular Systems	66
Kevin K. Anderson, Deanna L. Auberry, William R. Cannon* (William.Cannon@pnl.gov), Don S. Daly, Brian S. Hooker, Gregory B. Hurst, Jason E. McDermott, W. Hayes McDonald* (McDonaldWH@ornl.gov), Dale A. Pelletier, Denise D. Schmoyer* (SchmoyerDD@ornl.gov), Julia L. Sharp, Mudita Singhal* (Mudita.Singhal@pnl.gov), Ronald C. Taylor* (Ronald.Taylor@pnl.gov), Michelle V. Buchanan (BuchananMV@ornl.gov)	
41 Global Survey of Protein-Protein Interactions in <i>Rhodopseudomonas palustris</i>	67
Dale A. Pelletier* (pelletierda@ornl.gov), Gregory B. Hurst, Linda J. Foote, Trish K. Lankford, Catherine K. McKeown, Tse-Yuan S. Lu, Elizabeth T. Owens, Denise D. Schmoyer, Manesh B. Shah, Jennifer L. Morrell-Falvey, Brian S. Hooker, Stephen J. Kennel, W. Hayes McDonald, Mitchel J. Doktycz, Deanna L. Auberry, William R. Cannon, Kenneth J. Auberry, H. Steven Wiley, and Michelle V. Buchanan	
42 The Microbial Protein-Protein Interaction Database (MiPPI)	68
Denise D. Schmoyer* (schmoyerdd@ornl.gov), Sheryl A. Martin, Gregory B. Hurst, Manesh B. Shah, Dale A. Pelletier, W. Hayes McDonald, William R. Cannon, Deanna L. Auberry, and Michelle V. Buchanan	
43 Advances in Coverage and Quality for High-Throughput Protein-Protein Interaction Measurements	69
Jennifer Morrell-Falvey, Mitchel J. Doktycz, Dale A. Pelletier, Linda J. Foote, Elizabeth T. Owens, Sankar Venkatraman, W. Hayes McDonald* (mcdonaldwh@ornl.gov), Brian S. Hooker, Chiann-Tso Lin, Kristin D. Victry, Deanna L. Auberry, Eric A. Livesay, Daniel J. Orton, H. Steven Wiley, and Michelle V. Buchanan	
44 Genome-Wide Identification of Localized Protein Complexes in <i>Caulobacter</i>	71
P. Viollier* (patrick.viollier@case.edu), J. Werner, S. Pritchard, E. Chen, E. Huitema, L. Shapiro, and Z. Gitai (zgitai@princeton.edu)	

<u>Poster</u>	<u>Page</u>
45 The Structure and Function of the <i>Caulobacter</i> MreB Actin-Like Cytoskeleton	73
N. Dye,* M. Mielke, Z. Pincus, J. Theriot, L. Shapiro, and Z. Gitai (zgitai@princeton.edu)	
46 EM Tomography Enhancements	75
Fernando Amat* (famat@stanford.edu), Farshid Moussavi* (farshid1@stanford.edu), Kenneth H. Downing,, Mark Ellisman, Luis R. Comolli, Albert Lawrence, Mark Horowitz,* and Harley McAdams (hmcadams@stanford.edu)	
47 Automated Screen for Identification of Mislocalization and Morphological Mutants	76
G. Bowman, N. Hillson, M. Fero, S. Hong, L. Shapiro, and H. H. McAdams* (hmcadams@stanford.edu)	
48 Methods for <i>in vitro</i> and <i>in vivo</i> Imaging of Protein Complexes	78
Huilin Li* (hli@bnl.gov), James Hainfeld* (hainfeld@bnl.gov), Minghui Hu, Michael Mylenski, Kevin Ryan, Luping Qian, Raymond P. Briñas, Elena S. Lyman, and Larissa Kusnetsova	
49 Probing Single Microbial Proteins and Multi-Protein Complexes with Bioconjugated Quantum Dots	80
Gang Bao* (gang.bao@bme.gatech.edu), Grant Jensen, Shuming Nie, and Phil LeDuc	
50 Correlated Light and Electron Microscopy of Protein Complexes in <i>Caulobacter crescentus</i>	83
Guido M. Gaietta* (ggaietta@ncmir.ucsd.edu), Thomas J. Deerinck, Grant Bowman, Yi Chun Yeh, Luis R. Comolli, Lucy Shapiro, Harley McAdams, and Mark H. Ellisman	
51 Computational Analysis of the Protein Interaction Networks of Three Archaeal Microbes	84
Chris Ding* (chqding@lbl.gov), Chunlin Wang, and Stephen R. Holbrook	
52 The Use of Small Angle X-ray Scattering to Extract Low Resolution Structures and Monitor Sample Quality from Archeal Proteomes	85
Greg Hura* (glhura@lbl.gov), Michal Hammel, Susan Tsutakawa, Cesar Luna-Chavez, Robert Rambo, Ferris Poole, Francis Jenney, Angeli Lal Menon, Mike Adams, and John Tainer	
53 Imaging <i>Caulobacter crescentus</i> using Soft X-Ray Tomography: A New Imaging Tool for Genomics: GTL and Bioenergy Research	86
B.M. Maguire, C.A. Tonnessen, G. McDermott, A.J. McDonnell, M.A. Le Gros, and C.A. Larabell* (CALarabell@lbl.gov)	

GTL Milestone 2..... 89

Section 1: OMICS: Systems Measurements of Plants, Microbes, and Communities..... 89

54 The Virtual Institute of Microbial Stress and Survival: An Overview of the Environmental Stress Pathway Project..... 89

Carl Abulencia, Eric J. Alm, Gary Anderson, Edward Baidoo, Peter Benke, Sharon Borglin, Eoin L. Brodie, Romy Chakraborty, Swapnil Chhabra, Gabriela Chirica, Dylan Chivian, Michael J. Cipriano, M.E. Clark, Paramvir S. Dehal, Elliot C. Drury, Inna Dubchak, Dwayne A. Elias, Matthew W. Fields, J. Gabster, Sara P. Gaucher, Jil Geller, B. Giles, Masood Hadi, Terry C. Hazen, Qiang He, Zhili He, Christopher L. Hemme, E. Hendrickson, Kristina L. Hillesland, Hoi-Ying Holman, Katherine H. Huang, Y. Wayne Huang, C. Hwang, Janet Jacobsen, Marcin P. Joachimiak, Dominique C. Joyner, Jay D. Keasling, Keith Keller, Martin Keller, J. Leigh, T. Lie, Aindrila Mukhopadhyay, Richard Phan, Francesco Pingitore, Morgan Price, Alyssa M. Redding, Joseph A. Ringbauer Jr., Rajat Sapra, Christopher W. Schadt, Amy Shutkin, Anup K. Singh, David A. Stahl, Sergey M. Stolyar, Yinjie Tang, Joy D. Van Nostrand, Chris B. Walker, Judy D. Wall, Eleanor Wozei, Zamin K. Yang, Huei-Che Yen, Grant Zane, Aifen Zhou, Jizhong Zhou, and Adam P. Arkin* (aparkin@lbl.gov)

55 VIMSS ESPP Functional Genomics Core: Cell Wide Analysis of Metal-Reducing Bacteria .. 91

Aindrila Mukhopadhyay,* Edward Baidoo, Peter Benke, Swapnil Chhabra, Gabriela Chirica, Elliot Drury, Matthew Fields (fieldsmw@muohio.edu), Sara Gaucher, Masood Hadi, Qiang He, Zhili He, Chris Hemme, Jay Keasling (keasling@berkeley.edu), Francesco Pingitore, Alyssa Redding, Rajat Sapra, Anup Singh (aksingh@sandia.gov), Yinjie Tang, Judy Wall (wallj@missouri.edu), Huei-Che Yen, Grant Zane, Aifen Zhou, and Jizhong Zhou (jzhou@rccc.ou.edu)

56 Response of *Desulfovibrio vulgaris* Hildenborough to Acid pH..... 92

H.-C. Yen,* T. C. Hazen, Z. Yang, J. Zhou, K. H. Huang, E. J. Alm, A. P. Arkin (aparkin@lbl.gov), and J. D. Wall

57 Global Gene Regulation in *Desulfovibrio vulgaris* Hildenborough..... 93

Aifen Zhou,* Zhili He, Chris Hemme, Aindrila Mukhopadhyay, Jay Keasling, Adam P. Arkin (aparkin@lbl.gov), Terry C. Hazen, Judy D. Wall, and Jizhong Zhou

58 *Desulfovibrio vulgaris* Responses to Hexavalent Chromium at the Community, Population, and Cellular Levels..... 95

A. Klonowska,* Z. He, Q. He, M.E. Clark, S.B. Thieman, T.C. Hazen, E.L. Brodie, R. Chakraborty, E.J. Alm, B. Giles, H.-Y. Holman, A.P. Arkin (aparkin@lbl.gov), J.D. Wall, J. Zhou, and M.W. Fields

59 Energy Conserving Hydrogenases Drive Syntrophic Growth of *Desulfovibrio vulgaris* and *Methanococcus maripaludis* 96

C.B. Walker,* Z.K. Yang, Z. He, S.S. Stolyar, J. Jacobsen, J.A. Ringbauer Jr., J.D. Wall, J. Zhou, A. P. Arkin (aparkin@lbl.gov), and D.A. Stahl

60 A Large Number of Hypothetical Proteins are Differentially Expressed during Stress in *Desulfovibrio vulgaris* 97

Elliot C. Drury,* Alyssa M. Redding, Aindrila Mukhopadyay, Katherine H. Huang, Terry C. Hazen, Adam P. Arkin (aparkin@lbl.gov), Judy D. Wall, and Dwayne A. Elias*

<u>Poster</u>	<u>Page</u>
61 Phenotypic Correlations in <i>Desulfovibrio</i>	98
K.L. Hillesland,* C.B. Walker, and D.A. Stahl	
62 Nitrate Stress Response in <i>Desulfovibrio vulgaris</i> Hildenborough: Whole-Genome Transcriptomics and Proteomics Analyses	100
Qiang He,* Zhili He, Wenqiong Chen, Zamin Yang, Eric J. Alm, Katherine H. Huang, Huei-Che Yen, Dominique C. Joyner, Martin Keller, Adam P. Arkin (aparkin@lbl.gov), Terry C. Hazen, Judy D. Wall, and Jizhong Zhou	
63 Redox Proteomics in <i>Desulfovibrio vulgaris</i> Hildenborough: Search for Proteins that Mediate Stress Response via Post-Translational Modification of the Cys Residues	101
Rajat Sapra,* Sara Gaucher, Gabriela Chirica, Carrie Kozina, George Buffleben, Richard Phan, Dominique Joyner, Terry C. Hazen, Adam P. Arkin (aparkin@lbl.gov), and Anup K. Singh	
64 A Survey of Protein Post-Translational Modifications Found in the Sulfate-Reducing Bacterium <i>Desulfovibrio vulgaris</i> Hildenborough	103
Sara P. Gaucher,* Alyssa M. Redding, Gabriela S. Chirica, Rajat Sapra, George M. Buffleben, Carrie Kozina, Aindrita Mukhopadhyay, Dominique C. Joyner, Jay D. Keasling, Terry C. Hazen, Adam P. Arkin (aparkin@lbl.gov), David A. Stahl, Judy D. Wall, and Anup K. Singh	
65 The Ech Hydrogenase is Important for Growth of <i>D. vulgaris</i> with Hydrogen	104
S.M. Stolyar,* J. Wall, and D.A. Stahl	
66 Monitoring of Microbial Reduction and Reoxidation Activities in the FRC Sites using a Comprehensive Functional Gene Array	105
Zhili He,* Joy D. Van Nostrand, Liyou Wu, Terry J. Gentry, Ye Deng, Christopher W. Schadt, Weimin Wu, Jost Liebich, Song C. Chong, Baohua Gu, Phil Jardine, Craig Criddle, David Watson, Terry C. Hazen, and Jizhong Zhou	
67 Towards High-Throughput and High Sensitivity Approaches for Uncovering Total Environmental Gene Expression Patterns	107
Zamin Yang, Christopher W. Schadt,* Terry Hazen, and Martin Keller	
68 Experimental and Computational Approaches to Enhance Proteomics Measurements of Natural Microbial Communities	109
Nathan VerBerkmoes,* Mark Lefsrud, Chongle Pan, Brian Erickson, Manesh Shah, Chris Jeans, Steven Singer, Michael P. Thelen, Vincent Denef, I. Lo, Jillian Banfield (jill@eps.berkeley.edu), and Robert L. Hettich	
69 Strain-Resolved Proteogenomics-Enabled Ecological Study of Natural Microbial Communities Associated with Acid Mine Drainage Formation	111
V.J. Denef* (vdenef@berkeley.edu), N.C. VerBerkmoes, P. Wilmes, M. Shah, D. Goltsman, I. Lo, G. DiBartolo, L. Kalnejais, B.J. Baker, G.W. Tyson, J.M. Eppley, E.A. Allen, R.L. Hettich, M.P. Thelen, and J.F. Banfield	
70 A Novel Iron Oxidase Isolated from an Extremophilic Microbial Community	113
Steven W. Singer,* Christopher Jeans, Jason Raymond, Adam Zemla, Nathan C. VerBerkmoes, Robert L. Hettich, Clara Chan, Jill Banfield, and Michael P. Thelen (mthelen@lbl.gov)	

<u>Poster</u>	<u>Page</u>
71 Functional Analysis of Protein Phosphorylation in <i>Shewanella oneidensis</i> MR-1.....	114
C. Giometti* (csgiometti@anl.gov), G. Babnigg, A. Beliaev, G. Pinchuk, M. Romine, and J. Fredrickson	
72 Enriching Metabolic Function Predictions for <i>Shewanella oneidensis</i> MR-1 with Growth and Expression Studies	116
Margrethe H. Serres* (mserres@mbi.edu), Margaret F. Romine, Mike E. Driscoll, Tim S. Gardner, Natalia Maltsev, Miriam Land, Andrei Osterman, Mary Lipton, and LeeAnn McCue	
73 Proteomics Technologies Advance the Understanding of Microbial Systems Allowing for In-Depth Characterization of Microbes Important for Bioenergy Production, Bioremediation and Carbon Sequestration and Cycling.....	117
Mary S. Lipton* (mary.lipton@pnl.gov), Joshua Turse, Stephen Callister, Kim K. Hixson, Xuixia Du, Angela Norbeck, Samuel Purvine, Feng Yang, Margie F. Romine, Carrie D. Nicora, Joshua Adkins, Richard D. Smith, and Jim K. Fredrickson	
74 Functional Genomic Analysis of Current Production in High Power Density Microbial Fuel Cells.....	120
Kelly P. Nevin* (knevin@microbio.umass.edu), Sean F. Covalla, Jessica P. Johnson, Trevor L. Woodard, Raymond DiDonato Jr., Kim K. Hixson, Mary Lipton, and Derek R. Lovley	
75 Genome-Scale Analysis of Adaptive Evolution of <i>Geobacter</i> for Improved Metal Reduction and Electricity Production	122
Zarath Summers,* Kelly Nevin, Chris Herring, Richard Glaven, Shelley Haveman, James Elkins, Bernhard Palsson, and Derek Lovley (dlovley@microbio.umass.edu)	
76 Proteomic Profiling of the <i>Caulobacter crescentus</i> Cell Cycle and Starvation Response.....	123
Esteban Toro* (etoro@stanford.edu), Leticia Britos, Samuel O. Purvine, Mary S. Lipton, Tom Taverner, Feng Yang, Harley H. McAdams, Richard D. Smith, and Lucy Shapiro*	
77 Quantitative Shotgun Proteomics with ProRata: Application to Anaerobic Aromatic Degradation in <i>Rhodopseudomonas palustris</i>	124
C. Pan,* G. Kora, Y. Oda, D. Pelletier, N. C. VerBerkmoes, W. H. McDonald, G. Hurst, C. S. Harwood, R. L. Hettich (hettichrl@ornl.gov), and N. F. Samatova (samatovan@ornl.gov)	
78 From Genome to Metabolome: Correlating a System-Wide Response to Environmental Adaptation in a Hyperthermophile	126
Sunia A. Trauger* (strauger@scripps.edu), Ewa Kalisiak, Jarek Kalisiak, Hiro Morita, Angeli Lal Menon, Michael V. Weinberg, Farris L. Poole, Michael W. W. Adams, and Gary Siuzdak	
79 High Throughput Comprehensive and Quantitative Microbial and Community Proteomics.....	128
Richard D. Smith* (rds@pnl.gov), Joshua N. Adkins, David J. Anderson, Kenneth J. Auberry, Mikhail E. Belov, Stephen J. Callister, Therese R.W. Clauss, Jim K. Fredrickson, Xuixia Du, Kim K. Hixson, Navdeep Jaitly, Gary R. Kiebel, Mary S. Lipton, Eric A. Livesay, Anoop Mayampurath, Matthew E. Monroe, Ronald J. Moore, Heather M. Mottaz, Carrie D. Nicora, Angela D. Norbeck, Daniel J. Orton, Ljiljana Paša-Tolić, Kostantinos Petritis, David C. Prior, Samuel O. Purvine, Yufeng Shen, Anil K. Shukla, Aleksey V. Tolmachev, Nikola Tolić, Harold R. Udseth, Rui Zhang, and Rui Zhao	

<u>Poster</u>	<u>Page</u>
80 Exploring the Genome and Proteome of <i>Desulfitobacterium hafniense</i> DCB-2 for its Protein Complexes Involved in the Reduction of Selenium and Iron	130
Christina Harzman, Christi Hemming, Sang-Hoon Kim, David DeWitt, John Davis, Rachel Udelhoven, Kaitlin Duschene, Joan B. Broderick, James M. Tiedje, Terence L. Marsh* (marsht@msu.edu)	
Section 2: Metabolic Network Experimentation and Modeling	131
81 Improving the Production of Biotherapeutics using Metabolic Engineering	132
M. Bauman,* J. Jones, S. Krag, V. Ciccarone, D. Judd, S. Gorfien, Y.C. Lee, N. Tomiya, and M. Betenbaugh* (beten@jhu.edu)	
82 Improved Microbial Hydrogen Production by the Engineering of Specific Metabolic Segments of <i>Escherichia coli</i>	133
Zhanmin Fan, Ling Yuan, Yu Wang, and Ranjini Chatterjee* (rchatterjee@farasis.com)	
83 Toward the Automatic Generation of Genome-Scale Metabolic Models in the SEED	134
Matthew DeJongh* (dejongh@hope.edu) and Aaron Best	
84 Metabolic Engineering of Light and Dark Biochemical Pathways in Wild-Type and Mutant <i>Synechocystis</i> PCC 6803 Strains for Maximal, 24-Hour Production of Hydrogen Gas	135
P. S. Schrader, E. H. Burrows, F. W. R. Chaplen, and R. L. Ely* (ely@engr.orst.edu)	
85 Pathway Tools + MetaCyc = Comprehensive Pathway Modeling	137
Ron Caspi, Hartmut Foerster, Carol Fulcher, Michelle Green, Pallavi Kaipa, Markus Krummenacker, Mario Latendresse, Suzanne Paley, Chris Tissier, Peifen Zhang, Sue Rhee, and Peter D. Karp* (pkarp@ai.sri.com)	
86 Constraint-Based Modeling of Central Metabolism in the Family <i>Geobacteraceae</i>	139
Jun Sun* (jsun@genomatica.com), Steve Van Dien, Radhakrishnan Mahadevan, Maddalena Coppi, Laurie DiDonato, Carla Risso, Mounir Izallalen, Bradley Postier, Raymond DiDonato, Kai Zhuang, Priti Pharkya, Tom Fahland, Olivia Bui, Iman Famili, Christophe Schilling, and Derek Lovley	
87 Analysis of Degree of Genetic Redundancy in Prokaryotic Metabolic Networks	141
R. Mahadevan* (mahadevan@chem-eng.utoronto.ca) and D.R. Lovley	
88 Mechanisms of Sulfur Reduction by <i>Shewanella</i>	142
Edward J. Crane III* (EJ.Crane@pomona.edu), Evan T. Hall, and Ken Nealson	
89 Carbon and Energy Metabolism Strategies in <i>Shewanella</i>	143
G. Pinchuk* (Grigoriy.Pinchuk@pnl.gov), A. Beliaev, O. Geydebrekht, D. Kennedy, I. Famili, J. Reed, J. Scott, S. Reed, M. Romine, and J. Fredrickson	
90 Metabolic Reconstruction of <i>Shewanella oneidensis</i>: A Community Resource	145
Jennifer L. Reed, Iman Famili, Sharon J. Wiback, Christophe H. Schilling, Grigoriy Pinchuk, Margaret R. Romine, Johannes C. Scholten* (johannes.scholten@pnl.gov), Joel Klappenbach, and James K. Fredrickson	

<u>Poster</u>	<u>Page</u>
91 The Challenge of Incorporating Regulatory Effect in Genome-Scale Networks	146
C.L. Barrett*, M.J. Herrgard*, B.K. Cho, E.M. Knight, J. Elkins, and B.O. Palsson (palsson@ucsd.edu)	
92 Acclimation of <i>Chlamydomonas reinhardtii</i> to Anoxic Conditions: Gene Expression, Hydrogenase Induction and Metabolic Pathways	146
Michael Seibert* (mike_seibert@nrel.gov), Florence Mus, Alexandra Dubini, Maria L. Ghirardi, Matthew C. Posewitz, and Arthur R. Grossman	
93 Perspectives in Metabolic Flux Mapping	148
Jacqueline V. Shanks* (jshanks@iastate.edu)	
94 High-Resolution Functional Assignments of Genes through Mapping KEGG Pathways to Bacterial Genomes	149
Fenglou Mao, Hongwei Wu, and Ying Xu* (xyn@bmb.uga.edu)	
Section 3: Regulatory Processes	151
95 A Systems Approach to Characterizing Evolutionarily Conserved Transcriptional Complexes Elucidates the Architecture of a Global Regulatory Network in Archaea	151
Marc T. Facciotti*, David J. Reiss, Min Pan, Amardeep Kaur, Madhavi Vuthoori, Richard Bonneau, Paul Shannon, Alok Srivastava, Samuel M. Donahoe, Leroy Hood, and Nitin S. Baliga (nbaliga@systemsbiology.org)	
96 CRP and cAMP Regulatory Networks of <i>Shewanella oneidensis</i> MR-1 Involved in Anaerobic Energy Metabolism	152
Daad A. Saffarini*, Sheetal Shirodkar, Yang Zhang, and Alexander S. Beliaev	
97 Mapping the Genome-Scale Regulatory Network of <i>Shewanella oneidensis</i> MR-1: Identification of Metal-Respiratory Regulation	153
M.E. Driscoll, F.S. Juhn, J.J. Faith, B. Hayete, J.J. Collins, and T.S. Gardner* (tgardner@bu.edu)	
98 A Web-Based Tool for Visualizing <i>Shewanella</i> Gene Expression Profiles in Their Chromosomal Context	154
J.J. Faith*, R. Sachidanandam, and T.S. Gardner (tgardner@bu.edu)	
99 Comparative Genomics of Signal Transduction in <i>Shewanella</i>	155
Luke E. Ulrich and Igor B. Zhulin* (joulineib@ornl.gov)	
100 Comparative Genomics of Transcriptional Regulation of Metabolic Pathways in <i>Shewanella</i> Species	156
Dmitry Rodionov* (rodionov@burnham.org), Mikhail Gelfand, Margaret Romine, and Andrei Osterman	
101 Biological Aspects of Deciphering and Engineering Regulatory Networks	157
George N. Bennett* (gbennett@bioc.rice.edu)	

<u>Poster</u>	<u>Page</u>
102 Characterization of Behavioral Responses in <i>Shewanella oneidensis</i>	158
Jun Li, Margie Romine, and Mandy Ward* (mjward@jhu.edu)	
103 Development of <i>in vitro</i> Transcription System using Recombinant <i>Shewanella oneidensis</i> RNA Polymerase	160
Younggyu Kim* (ykim@chem.ucla.edu), Sam On Ho, Natalie Gassman, and Shimon Weiss (sweiss@chem.ucla.edu)	
104 Genetic Analysis of Anaerobic Respiration in <i>Shewanella oneidensis</i> MR-1	162
Jizhong Zhou, Haichun Gao,* Xiaohu Wang, Soumitra Barua, Yunfeng Yang, Samantha B. Reed, Dave Culley, Zamin Yang, Christopher Hemme, Zhili He, Margaret Romine, Kenneth Nealson, James M. Tiejde, Timothy Palzkill, and James K. Fredrickson	
105 A Phylogenetic Gibbs Sampler for High-Resolution Comparative Genomics Studies of Transcription Regulation	164
William A. Thompson, Sean P. Conlan, Thomas M. Smith, Lee A. Newberg, Lee Ann McCue* (leeann.mccue@pnl.gov), and Charles E. Lawrence	
106 Challenges in Predictive Modeling for Engineering/Deciphering the Regulatory Networks	165
James C. Liao*	
107 Molecular Mechanisms Regulating Gene Expression in <i>Geobacter sulfurreducens</i> under Environmentally Relevant Conditions	165
Toshiyuki Ueki* (tueki@microbio.umass.edu), Ching Leang, Byoung-Chan Kim, Richard Glaven, Haiping Ke, Katy Juárez, and Derek R. Lovley (dlovley@microbio.umass.edu)	
108 Computational Analysis of Transcription Regulation of <i>Geobacter sulfurreducens</i>	167
Julia Krushkal* (jkrushka@utmem.edu), Marko Puljic, Ronald M. Adkins, Jeanette Peeples, Bin Yan, Ching Leang, Laurie N. DiDonato, Cinthia E. Núñez, Toshiyuki Ueki, Radhakrishnan Mahadevan, Brad Postier, Barbara Methé, and Derek R. Lovley (dlovley@microbio.umass.edu)	
109 Identification of Small Non-Coding RNAs and Acceptance Rate Studies in Members of the <i>Geobacteraceae</i>	170
Barbara Methé* (bmethe@tigr.org), Robert DeBoy, Sean Daugherty, Ty Arrington, Kelly Nevin, Jonathan Badger, and Derek Lovley	
110 Bacterial Cell Cycle Control System and a Control System Simulation Model	171
J. Collier, X. Shen, M. Horowitz, L. Shapiro, and H. H. McAdams* (hmcadams@stanford.edu)	
111 Automated Accurate, Concise and Consistent Product Description Assignment for Microbial Regulatory Proteins	174
Loren Hauser* (hauserlj@ornl.gov), Frank Larimer, and Miriam Land	

<u>Poster</u>	<u>Page</u>
GTL Milestone 3	175
Section 1: Computing Infrastructure, Bioinformatics, and Data Management	175
112 Center for Computational Biology at the University of California, Merced	175
Michael Colvin* (mcolvin@ucmerced.edu), Arnold Kim, Masa Watanabe,* and Felice Lightstone	
113 Projects from the DOE-BACTER Institute at the University of Wisconsin, Madison	176
Julie C. Mitchell* (mitchell@math.wisc.edu), Julie Simons, Paul Milewski, Peter Koenig, and Qiang Cui	
114 VIMSS Computational Core	177
Paramvir S. Dehal* (PSDehal@lbl.gov), Eric J. Alm, Dylan Chivian, Katherine H. Huang, Y. Wayne Huang, Janet Jacobsen, Marcin P. Joachimiak, Keith Keller, Morgan N. Price, and Adam P. Arkin (aparkin@lbl.gov)	
115 RegTransBase – A Resource for Studying Regulatory Interactions and Regulon Predictions in Bacteria	179
Michael J. Cipriano,* Alexei E. Kazakov, Dmitry Ravcheev, Adam P. Arkin (aparkin@lbl.gov), Mikhail S. Gelfand, and Inna Dubchak	
116 MicrobesOnline: An Integrated Portal for Comparative Functional Genomics	180
Marcin P. Joachimiak,* Katherine H. Huang, Eric J. Alm, Dylan Chivian, Paramvir S. Dehal, Y. Wayne Huang, Janet Jacobsen, Keith Keller, Morgan N. Price, and Adam P. Arkin (aparkin@lbl.gov)	
117 Protein Complex Analysis Project (PCAP): Data Management and Bioinformatics Subproject	181
Adam P. Arkin, Ralph Santos, Y. Wayne Huang, Janet Jacobsen, Keith Keller, Steven S. Andrews, Steven E. Brenner, Max Shatsky, and John-Marc Chandonia* (JMChandonia@lbl.gov)	
118 Gaggles: A Framework for Database Integration and Software Interoperability	182
Christopher Bare, Paul Shannon, Michael Johnson, and Nitin S. Baliga* (nbaliga@systemsbiology.org)	
119 Sensitivity Analysis on MS2 Viral Dynamics Using Interval Mathematics	183
Ozlem Yilmaz,* Luke E. K. Achenie (achenie@engr.uconn.edu), and Ranjan Srivastava	
120 The BioWarehouse System for Integration of Bioinformatics Databases	184
Tom Lee, Valerie Wagner, Yannick Pouliot, and Peter D. Karp* (pkarp@ai.sri.com)	
121 A Cell Centered Database for Microbial Cells	185
Maryann E. Martone,* Joy Sargis, Andrew McDonnell, Gerry McDermott, Carolyn Larabell, Mark Le Gros, Joshua Tran, Willy Wong, Vincent Ye, Harley McAdams (hmcadams@stanford.edu), and Mark H. Ellisman	
122 Developments in the Systems Biology Workbench	186
Frank Bergmann,* Anastasia Deckard, and Herbert M. Sauro (hsauro@u.washington.edu)	

<u>Poster</u>	<u>Page</u>
123 The Ribosomal Database Project II: Introducing <i>myRDP</i> Space and Quality-Controlled Public Data	188
J. R. Cole* (colej@msu.edu), Q. Wang, B. Chai, E. Cardenas, R. J. Farris, A. M. Bandela, A. S. Kulam-Syed-Mohideen, D. M. McGarrell, G. M. Garrity, and J. M. Tiedje	
124 An Integrated Knowledge Resource for the <i>Shewanella</i> Federation	190
Nagiza F. Samatova* (samatovan@ornl.gov), Denise Schmoyer, Tatiana Karpinets, Guruprasad Kora, Sergey Passovets, Michael Leuze, and Ed Uberbacher (ube@ornl.gov) Collaborators from Shewanella Federation: Timothy S. Gardner, Gyorgy Babnigg, Carol S. Giometti, Margrethe Serres, Anna Obratzsova, Grigoriy E. Pinchuk, Alexander Beliaev, Margaret F. Romine, Kenneth Nealson, and James K. Fredrickson	
125 Informatics Strategies for Large-Scale Novel Cross-linker Analysis	193
Gordon A. Anderson* (Gordon@pnl.gov) , Nikola Tolic, Xiaoting Tang, and James E. Bruce	
Communication	195
126 Communicating Genomics:GTL	195
Anne E. Adamson, Shirley H. Andrews, Jennifer L. Bownas, Sherry A. Estes, Sheryl A. Martin, Marissa D. Mills, Kim Nylander, Judy M. Wyrick, Anita J. Alton, and Betty K. Mansfield* (mansfieldbk@ornl.gov)	
Ethical, Legal, and Societal Issues	197
127 Science Literacy Project for Public Radio Journalists	197
Bari Scott* (bariscot@aol.com)	
128 The DNA Files®	199
Bari Scott* (bariscot@aol.com)	
USDA and DOE Joint Research	201
129 Manipulation of Lignin Biosynthesis to Maximize Ethanol Production from <i>Populus</i> Feedstocks	201
Clint Chapple (chapple@purdue.edu), Richard Meilan, and Michael Ladisch	
130 Systematic Modification of Monolignol Pathway Gene Expression for Improved Lignocellulose Utilization	202
Richard Dixon (PI)(radixon@noble.org) and Fang Chen	
131 Sorghum Biomass/Feedstock Genomics Research for Bioenergy	203
William Rooney (wlr@tamu.edu), John Mullet, Steve Kresovich, Doreen Ware, and P. Klein	
132 Streamlined Method for Biomass Whole-Cell-Wall Structural Profiling	204
John Ralph (jralph@wisc.edu), F. Lu, B. Sundberg, and S. Mansfield	

<u>Poster</u>	<u>Page</u>
133 Development of a Proteoglycan Chip for Plant Glycomics	205
Chris R. Somerville (crs@stanford.edu)	
134 Biochemical Genomics of Wood Formation: O-Acyltransferases for Alteration of Lignocellulosic Property and Enhancement of Carbon Deposition in Poplar	206
Chang-Jun Liu (cliu@bnl.gov)	
135 Genomic Knowledgebase for Facilitating the Use of Woody Biomass for Fuel Ethanol Production	207
Vincent L. Chiang (vincent_chiang@ncsu.edu)	
136 Genetic Dissection of the Lignocellulosic Pathway of Wheat to Improve Biomass Quality of Grasses as a Feedstock for Biofuels	208
Bikram Gill (bsgill@ksu.edu) and Wanlong Li	
137 Using Association Mapping to Identify Markers for Cell Wall Constituents and Biomass Yield in Alfalfa	209
Charles Brummer (brummer@uga.edu), Kenneth J. Moore, and Jeff J. Doyle	
Appendix 1: Participants	211
Appendix 2: Web Sites	219
Author Index.....	221
Institutional Index.....	226

Workshop Abstracts

The Genomics:GTL Program abstracts and posters are organized according to the Milestones set forth in the GTL Roadmap and as shown below [*DOE Genomics:GTL Roadmap: Systems Biology for Energy and Environment*; October 2005; GenomicsGTL.energy.gov]. Some of the research projects are pilots or proof-of-principle studies for systems biology, technology and methods development, and computing.

Abstracts associated with the Metabolic Engineering Working Group (MEWG) are identified as such and intermixed with GTL abstracts in relevant categories.

Abstracts from the Plant Feedstock Genomics Joint Program (PFGJG) are in their own subsection beginning on page 201.

Genomics:GTL Overarching Scientific Goal

Achieve a predictive, systems-level understanding of biological systems to help enable biobased solutions to DOE mission challenges.

Science and Technology Milestones

Milestone 1: Develop Techniques to Determine the Genome Structure and Functional Potential of Microbes and Microbial Communities

- 1.1 Organism Sequencing, Annotation, and Comparative Genomics
- 1.2 Microbial-Community Sequencing and Analysis
- 1.3 Protein Production and Characterization
- 1.4 Molecular Interactions

Milestone 2: Develop Methods and Concepts Needed to Achieve a Systems-Level Understanding of Microbial Cell and Community Function, Regulation, and Dynamics

- 2.1 Omics: Systems Measurements of Plants, Microbes, and Communities
- 2.2 Metabolic Network Experimentation and Modeling
- 2.3 Regulatory Processes

Milestone 3: Develop the Knowledgebase, Computational Methods, and Capabilities to Advance the Understanding and Prediction of Complex Biological Systems

- 3.1 Computing Infrastructure, Bioinformatics, and Data Management

Communication

Ethical, Legal, and Societal Issues

USDA-DOE Abstracts

The following table is a simple summation of how GTL science and DOE missions align (GTL Roadmap p. 40).

Summary Table. GTL Science Roadmap for DOE Missions			
DOE Mission Goals		GTL Science Roadmaps	
Selected Processes	Biofuels	Processes to convert cellulose to fuels - Understanding and improving cellulase activity - Improving sugar transportation and fermentation to alcohols - Integrated processing Microbial processes to convert sunlight to hydrogen fuels - Understanding photolytic fuel production - Designing photosynthetic biofuel systems	Science Objectives ► Characterize genes, proteins, machines, pathways, and systems - Conducting genomic surveys and comparisons - Mining natural systems for new functions - Producing and characterizing proteins - Analyzing interactions, complexes, and machines ► Understand functions and regulation - Measuring molecular responses: Inventories - Performing functional assays ► Develop predictive mechanistic models - Conducting experimental design - Designing and manipulating molecules - Using cellular and cell-free systems
	Environmental Remediation	Microbial processes to reduce toxic metals - Understanding microbe-mineral interactions - Devising restoration processes	Mission Outputs Systems engineering - System-design strategies for deployment - Living and extracellular systems - Validation and verification analyses
Natural Systems' Behavior	Carbon Cycling and Sequestration	Subsurface microbial communities' role in transport and fate of contaminants - Understanding fate and effects - Supporting remediation decisions Ocean microbial communities' role in the biological CO₂ pump - Understanding C, N, P, O, and S cycles - Predicting climate responses - Assessing impacts of sequestration Terrestrial microbial communities' role in global carbon cycle - Understanding C, N, P, O, and S cycles - Predicting carbon inventories and climate responses - Assessing sequestration concepts	Science Objectives ► Analyze communities and their genomic potential - Sequencing and comparing genomes - Screening natural systems for processes - Producing and characterizing proteins ► Understand community responses, regulation - Comparing CO₂, nutrients, biogeochemistry cycles - Producing cellular and community molecular inventories - Performing community functional assays ► Predict responses and impacts - Building interactive and predictive models - Applying natural and manipulated scenarios
			Mission Outputs Robust science base for policy and engineering - Model ecosystem response to natural events - Efficacy and impacts of intervention strategies Sensor development - Community dynamics - Environmental and functional assays
A capsule summary of systems being studied, mission goals that drive the analysis, generalized science roadmaps, and outputs to DOE missions. To elucidate design principles, each of these goals entails the examination of thousands of natural primary and ancillary pathways, variants, and functions, as well as large numbers of experimental mutations.			

GTL Milestone 1

Develop Techniques to Determine the Genome Structure and Functional Potential of Microbes and Microbial Communities

Section 1

Organism Sequencing, Annotation, and Comparative Genomics

1 ^{GTL}

Genomic Reconstruction and Experimental Validation of Catabolic Pathways in *Shewanella* Species

Andrei Osterman^{1,2*} (osterman@burnham.org), Dmitry Rodionov,¹ Chen Yang,¹ Yanbing Wang,⁴ Margaret Romine,³ Anna Obraztsova,^{4*} and **Kenneth Nealson⁴**

¹Burnham Institute for Medical Research, La Jolla, California; ²Fellowship for Interpretation of Genomes, Burr Ridge, Illinois; ³Pacific Northwest National Laboratory, Richland, Washington; and ⁴Department of Earth Sciences, University of Southern California, Los Angeles, California

Project Goals: This project is a component of the *Shewanella* Federation and as such contributes to the overall goal of applying the tools of genomics, leveraging the availability of genome sequence for 18 additional strains of *Shewanella*, to better understand the ecophysiology and speciation of respiratory-versatile members of this important genus. To understand these systems the SF is using genome-based approaches to investigate *Shewanella* as a system of integrated networks; first describing key cellular subsystems—those involved in signal transduction, regulation, and metabolism—then building towards understanding the function of whole cells and, eventually, cells within populations. As a general approach, the SF is collectively employing complimentary “top-down”—bioinformatics-based genome functional predictions, high-throughput expression analyses, and functional genomics approaches to uncover key genes as well as metabolic and regulatory networks. The “bottom-up” component employs more traditional approaches including genetics, physiology and biochemistry to test or verify predictions. This information will ultimately be linked to analyses of signal transduction and transcriptional regulatory systems and used to develop a linked model that will contribute to understanding the ecophysiology of *Shewanella* in redox stratified environments.

Knowledge of the pathways and mechanisms of carbon assimilation and utilization for biomass and energy production is one of the key aspects of our understanding of environmental microorganisms in the context of respective ecosystems. Comparative genomic analysis has revolutionized our ability to quickly predict which metabolic subsystems occur in newly sequenced genomes, the set of genes of which each is comprised, and to suggest their functional roles within each subsystem. Results of this predictive analysis can then be used to design and conduct targeted physiological and biochemical assays to validate novel conjectures of gene function revealed in this process. By taking advantage of such computational predictions one can dramatically reduce the volume of experimental studies required to assess basic metabolic properties of multiple bacterial species.

The current availability of partial or complete genome sequences for 16 *Shewanella* strains provides and unprecedented opportunity for a systematic comparative analysis of this important group of species. For example, the integrative genomic approach was successfully used in our recent analysis of the novel N-acetylglucosamine (GlcNAc) utilization subsystem in *S. oneidensis* and related species. By using subsystem reconstruction and genome context analysis tools provided in The SEED genomic platform (<http://theseed.uchicago.edu/FIG/index.cgi>) we tentatively assigned a number of novel genes including GlcNAc-related transporter (NagP, SO3503), transcriptional regulator (NagR, SO3516) and enzymes, GlcNAc kinase (SO3507) and GlcN-6P deaminase (SO3506) that are non-homologous to the respective components of *E. coli* GlcNAc pathway. Two latter assignments and the whole biochemical pathway of GlcNAc conversion to Fructose-6P were experimentally verified by in vitro reconstitution. The results of phenotypic profiling were fully consistent with genomic reconstruction as only one strain, *S. frigidimarina*, which lacked the respective genes, was unable to grow on GlcNAc.

Extension of this approach to the analysis of a large panel of catabolic pathways has revealed substantial differences between *Shewanellae* and well-studied model species, such as *E. coli*. These differences occur at the level of presence or absence of the entire pathways, the use of alternative biochemical routes, different regulatory mechanisms and nonorthologous displacements of individual genes. For example, in contrast to *E. coli*, all of the sequenced *Shewanella* ssp. possess the elaborate machinery for catabolism of branch chain amino acids and histidine. Under experimental conditions used so far, most (but not all) of the tested *Shewanellae* were unable to grow on these amino acids as sole carbon sources. Further experiments are currently under way to reconcile these apparent inconsistencies. At the same time, both bioinformatics and experimental analyses revealed a fully consistent distribution of glycerate and sucrose utilization pathways among different strains of *Shewanella*. The genomic reconstruction of both pathways (not present in *E. coli*) allowed U.S. to predict novel gene families, including putative glycerate transporter (e.g., SO1771) as well as regulatory and uptake components of sucrose metabolism (in the *scr* operon, Sfri1095-1099, of *S. frigidimarina*). Using a genome context, such as clustering on the chromosome and long-range homology analysis, we were able to predict candidate genes for various functional roles in other pathways (e.g., fatty acid metabolism). In addition to the physiological analysis of the *Shewanellae*, we are currently pursuing experimental validation of several functional predictions by a variety of biochemical and genetic techniques. Results of these analyses are an integral component of ongoing research of *Shewanella* **Federation** whose ultimate goal is to develop a better understanding of *Shewanella* ecophysiology and speciation.

2 ^{GT}

Evolutionary Analysis of Proteins Deduced from 10 Fully Sequenced *Shewanella* Genomes

N. Maltsev^{1*} (maltsev@mcs.anl.gov), D. Sulakhe,¹ A. Rodriguez,¹ M. Syed,¹ and M. Romine²

¹Argonne National Laboratory, Argonne, Illinois and ²Pacific Northwest National Laboratory, Richland, Washington

Project Goals: The proteomes of the 10 completed *Shewanella* genomes (total 43,839 protein sequences) were analyzed using GNARE (GeNome Analysis and Research Environment). This Grid-based computational system for automated high-throughput analysis of genomes and metabolic reconstructions is being developed by Bioinformatics group at MCS, Argonne National

Laboratory. In addition to offering users with tools for annotating protein functions, GNARE provides automated metabolic reconstructions from the sequence data to facilitate identification of missing enzymes and comparative analysis of metabolic pathways. GNARE allows comparison of metabolic models of *Shewanella* strains with over 300 metabolic models for completely sequenced prokaryotic organisms. Chisel, a workbench for evolutionary analysis of enzymes being developed by our group, was used as a supporting tool for automated prediction of function, identification of taxonomy-specific metabolic signatures and identification of cases of potential horizontal gene transfer. The poster will present a detailed examination of these analyses.

The Joint Genome Institute has produced closed genome sequences for nine *Shewanella* strains. The environments from which these 9 strains and the first strain sequenced, *S. oneidensis* MR-1, were collected from vary from fresh and marine waters and underlying sediments to terrestrial sediments. The availability of this diverse set of genomes provides a unique opportunity to explore protein and metabolic pathway evolution within a single Genus. We will provide an overview of the unique suite of computational tools that we have begun applying to this set of sequences for both the purposes of updating the annotation and for studying cellular evolution.

Genome Annotation Using GNARE. The proteins deduced from the 10 completed *Shewanella* genomes (total 43,839 protein sequences) were analyzed using GNARE (GeNome Analysis and Research Environment)—a Grid-based computational system for automated high-throughput sequence characterization and metabolic reconstructions. This system uses conventional bioinformatics tools including BLAST, Blocks, InterPro as well as specialty tools developed by our group, including Chisel and a function prediction voting algorithm. The analysis was performed on 1317 CPUs of distributed computational resources from the Open Science Grid and TeraGrid allowing analysis of all 10 *Shewanella* proteomes to be completed in 64 hours. The process yields a suite of web pages that enables the *Shewanella* federation annotation team to view the results of the automated analysis and that facilitates further searches for evidence of function via a broad variety of additional computational tools provided by Puma2. Because the system was specifically designed for cross-genome comparative analysis it is well suited for comparative analysis of predicted functions associated with the various *Shewanella* proteomes.

Metabolic reconstructions. In addition to providing users with tools for annotating protein functions, GNARE provides automated metabolic reconstructions (accessible via a web interface) for each genome that can be viewed via either KEGG or EMP maps to facilitate identification of missing enzymes and subsequent searches for candidate proteins to fill these metabolic pathway gaps through use of high-resolution tools, such as Dragonfly and PhyloBlocks, developed by our group. As annotations are updated in the GNARE system they can readily be propagated through the metabolic constructions, thereby further facilitating the process of predicting protein functions in context of metabolic pathways. Simultaneous analysis of all 10 proteomes provides a foundation for comparative analysis of metabolic pathways characteristic of particular strains. We will present examples of the resulting predicted metabolic diversity among the various strains of *Shewanella*.

Identification of taxonomy-specific metabolic signatures using Chisel. A workbench for identification and evolutionary analysis of taxonomic and phenotypic variations of enzymes Chisel was used as a supporting tool for automated prediction of function for the 10 *Shewanella* proteomes. Chisel predicted enzyme assignments for 9,946 of a total 43,691 proteins in the 10 genomes. The number of EC assignments for individual genomes ranged from 718 in *S. denitrificans* OS217 to 1,081 for *S. oneidensis* MR-1. Interestingly, out of 1,081 enzymes predicted by Chisel in MR-1, 598 proteins corresponded to enzyme variations specific to *Proteobacteria*; 217 proteins to *Gammaproteobacteria*; 11 to *Alteromonadales* and 39 enzymes to the *Shewanella* Genus. Such variations in levels of taxonomic specificity indicate that enzymes in metabolic pathways evolve at different rates. In the course of

adaptation some of the enzymes become more specific to particular taxonomies. The *Shewanella*-specific variations of enzymes were found to be associated predominantly with core metabolic pathways (e.g., glycolysis, purine and pyrimidine biosynthesis, metabolism of amino acids) as well as chemotaxis and sensory transduction processes. This observation suggests significant systems-level adaptation that led to diversification of enzymes in this group of organisms in the course of evolution. Identification of taxonomy-specific signature enzymes may provide insights into mechanisms driving the emergence of taxonomy and phenotype-specific pathways. The Chisel system also supports the development of PCR primers and oligonucleotides corresponding to these models using the CODE-HOP program (Henikoff and Henikoff, 1996). This feature can assist experimentalists in identifying particular enzymatic functions in organisms of interest using biochip- or PCR-based technologies.

Discovery of potential cases of horizontal gene transfer. The Chisel analysis also helps to identify potential cases of horizontal gene transfer. According to our analysis a significant number of enzymatic genes in *Shewanella* appear to have been acquired from *Cyanobacteria* and *Firmicutes*. For example, analysis of the MR-1 revealed 18 proteins most similar to Cyanobacterial enzymes and 36 enzymes that are most closely related to various *Firmicutes*.

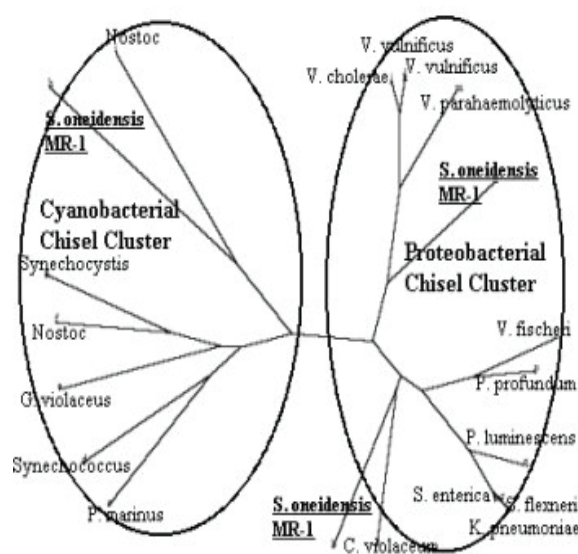


Fig 2. Phylogenetic tree for peptide deformylase (EC 3.5.1.88). MR-1 has 3 copies of this enzyme: 2 were classified by Chisel algorithm to a proteobacterial cluster (SF004749_6_B_Proteobacteri8) and one to a cyanobacterial cluster (SF004749_6_B_Cyanobacteria4).

The poster will present a detailed examination of these findings.

3 ^{GT}LModeling Conserved Indels as Phylogenetic Markers in *Shewanella*John P. McCrow^{1*} (mccrow@usc.edu), Kenneth H. Nealson,² and Michael S. Waterman¹¹Computational Molecular Biology, University of Southern California, Los Angeles, California and²Geobiology, University of Southern California, Los Angeles, California

Project Goals: This project is a component of the *Shewanella* Federation and as such contributes to the overall goal of applying the tools of genomics, leveraging the availability of genome sequence for 18 additional strains of *Shewanella*, to better understand the ecophysiology and speciation of respiratory-versatile members of this important genus. To understand these systems the SF is using genome-based approaches to investigate *Shewanella* as a system of integrated networks; first describing key cellular subsystems those involved in signal transduction, regulation, and metabolism—then building towards understanding the function of whole cells and, eventually, cells within populations. As a general approach, the SF is collectively employing complimentary “top-down” bioinformatics-based genome functional predictions, high-throughput expression analyses, and functional genomics approaches to uncover key genes as well as metabolic and regulatory networks. The “bottom-up” component employs more traditional approaches including genetics, physiology and biochemistry to test or verify predictions. This information will ultimately be linked to analyses of signal transduction and transcriptional regulatory systems and used to develop a linked model that will contribute to understanding the ecophysiology of *Shewanella* in redox stratified environments.

Phylogenetic signal derived from small subunit rRNA or core protein sequences is especially confusing and noisy among bacterial genomes, perhaps partially due to elevated levels of lateral gene transfer between lineages. The use of Rare Genomic Changes (RGCs) as phylogenetic characters has been proposed as potentially yielding higher quality information about the evolutionary history of genomes than collections of nucleotide or amino acid substitutions. High level genomic information including gene content, gene order, conserved inserts or deletions (indels), and nucleotide composition such as GC content, have been used to track evolutionary relationships between organisms within common core gene sets. However, the use of RGCs is limited in mainstream phylogenetic analysis because, unlike for conserved amino acid sequences, no statistical models exist to describe these high level genomic events.

We describe a method for the direct alignment, statistical modeling, and integration of conserved indels with adjacent amino acid sequence for improved phylogenetic signal. The best multiple alignment of amino acid sequences is not suitable for directly aligning conserved indels. Instead, we perform all pairwise alignments around potential indels, and filter out those with ambiguous classifications, as far as conserved indel states, to reduce homoplasy. The resulting set of conserved indels defines the highest confidence splits and are used as phylogenetic characters to infer the historical branching order of the species involved. While the extent of homoplasy within conserved indels is assumed to be low, we know of no previous statistical models or attempts to quantify independent convergent events or reversion rates of conserved indels. By integrating adjacent amino acid sequences with their corresponding conserved indels we can estimate the extent of homoplasy within conserved indels themselves, as well as improve the accuracy of phylogenetic signal from conserved proteins.

Conserved indels have great potential in aiding our understanding and validation of both the historical branching order of species, as well as the identification of lateral gene transfer events. They may

also be particularly suited for use as simple molecular markers of lineage and have the potential to be used to easily identify species present in an unknown mixed or pure sample without the need for any sequencing. Here we focus on exploring the utility of conserved indels to describe the phylogenetic relationships among sequenced species of the bacterial genus *Shewanella*.

4 ^{GTL}

***Shewanella* Population Comparative Genomics and Proteomics: Connecting Speciation, Ecophysiology, and Evolution**

Jorge L.M. Rodrigues^{1*} (rodrig76@msu.edu), Konstantinos T. Kostantinidis,² Margaret F. Romine,³ Margrethe H. Serres,⁴ Lee Ann McCue,³ Mary S. Lipton,³ Carol S. Giometti,⁴ Anna Obraztova,⁵ Matt Marshall,³ Miriam Land,⁶ Kenneth H. Nealson,⁵ James K. Fredrickson,² and **James M. Tiedje**¹

¹Michigan State University, East Lansing, Michigan; ²Massachusetts Institute of Technology, Cambridge, Massachusetts; ³Pacific Northwest National Laboratory, Richland, Washington; ⁴Marine Biological Laboratory, Woods Hole, Massachusetts; ⁵Argonne National Laboratory, Argonne, Illinois; ⁶University of Southern California, Los Angeles, California; and ⁶Oak Ridge National Laboratory, Oak Ridge, Tennessee

Project Goals: Integrated genome-based studies of *Shewanella* ecophysiology. The overall goal of this project is to apply the tools of genomics, to better understand the ecophysiology and speciation of respiratory-versatile members of the *Shewanella* genus.

Shewanella is a very versatile microorganism, capable of respiring more than 10 electron acceptors and found in many different environments. Therefore, *Shewanella* has potential to be used in different bioremediation schemes ranging from nitrate contaminant removal to metal and radionuclides reduction/immobilization processes. Previous molecular and physiological studies have primarily focused on *Shewanella oneidensis* MR-1, but it is still unclear whether prokaryotic systems biology can rely on a single model organism to predict functional responses for its entire population. Here, we extend the knowledge about the *Shewanella* genus by comparing closed genomes and their respective proteomes of 10 strains of this genus, including: *S. oneidensis* MR-1, *S. putrefaciens* W3-18-1, *S. putrefaciens* CN-32, *Shewanella* sp. MR-7, *Shewanella* sp. MR-4, *Shewanella* sp. ANA-3, *S. denitrificans* OS217, *S. frigidimarina* NCIMB 400, *S. loihica* PV-4, and *S. amazonensis* SB2B. The availability of these genomes allows questions towards the following aims: 1) to identify the gene core content of *Shewanella* genus, 2) to find genetic differences responsible for the ecological and physiological differentiation among the strains, and 3) to identify the ecological forces being implemented at genome level leading to speciation.

Results from comparative genomics showed that the above strains vary from 70 to 98.4% on pairwise average nucleotide identity (ANI). These results reveal a continuum of genetic relatedness for all sequenced genomes. The gene core dataset was calculated with use of three different methods giving similar values: 1817 for reciprocal DNA best match (Konstantinidis and Tiedje (2005), 1984 for protein alignments of 70% and a scoring matrix with Pam value of 100 (Serres and Riley 2006), and 2075 for pairwise reciprocal BLAST. The gene core dataset identified in all three sets is 1718 genes. Genome synteny deteriorated rapidly as the ANI decreased to values below 80%, indicating extensive chromosomal rearrangements that might have significant functional impact on the phenotypic and proteomic profiles of *Shewanella* species. The predicted central metabolism is almost identical and we have identified over 90 conserved pathways to this date for all sequenced genomes. A survey of

unique genes belonging to each of the strains revealed that the majority fell into select categories: 1) hypotheticals, 2) mobile elements, 3) motility and attachment, 4) sensory, and 5) regulatory genes. Co-localization of many of the unique genes on the genome suggests that many may have been acquired via lateral transfer. These differences might indicate that ecological forces are being implemented at the genome level, allowing short term niche adaptation (for closely related strains), leading to later speciation (distantly related species).

Proteomic analyses using two-dimensional electrophoresis and ion trap mass spectrometry were performed for all sequenced strains, resulting in larger differences at proteomic level in comparison to genomic analyses. These results might indicate extensive differences at the regulatory level, since all strains were grown under identical conditions.

Shewanella denitrificans is the only member of this group unable to reduce iron and hence provides an excellent tool to investigate which genes may be responsible for adaptation to this metabolic resource. Functions unique to the metal reduced that were highlighted by this type of comparative analysis include: 1) quinol:fumarate reductase, 2) Fe(III) permease, 3, glycogen metabolism, 4) lactate oxidation, 5) NiFe hydrogenase, and 6) a large cluster of fatty acid biosynthetic genes. This strain is devoid of any proteins containing more than four cytochrome c heme binding motifs or of metaquinone biosynthetic genes. While also found in *S. denitrificans* an alternative variant of Na-translocating NADH-quinone reductase and ammonifying nitrate reductase was found only in the metal reducing *Shewanellae*. *S. denitrificans* lacks various signal transduction/regulatory proteins and transporters that are believed to be associated with anaerobic metabolism.

These results highlight the power of comparative bioinformatics, proteomics, and phenotypic analyses of related sequenced strains to acquire a greater understanding of evolution and ecophysiology speciation.

References

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.
2. Serres, M.H. and M. Riley. 2006. 2006. Genomic analysis of carbon source metabolism of *Shewanella oneidensis* MR-1: predictions versus experiments. *J. Bacteriol.* **188**:4601-4609.

5 ^{GTL}

The Complete Genome of the Uncultivated Ultra-Deep Subsurface Bacterium *Desulforudis audaxviator* Obtained by Environmental Genomics

Dylan Chivian^{1,2*} (DCChivian@lbl.gov), Eric J. Alm,^{1,3} Eoin L. Brodie,² David E. Culley,⁴ Thomas M. Gihring,⁵ Alla Lapidus,⁶ Li-Hung Lin,⁷ Steve Lowry,⁶ Duane P. Moser,⁸ Paul Richardson,⁷ Gordon Southam,⁹ Greg Wanger,⁹ Lisa M. Pratt,¹⁰ **Adam P. Arkin**^{1,2,11,12,13} (aparkin@lbl.gov), Terry C. Hazen,^{1,2} Fred J. Brockman,⁴ and Tullis C. Onstott¹⁴

¹Virtual Institute for Microbial Stress and Survival, <http://vimss.lbl.gov>; ²Lawrence Berkeley National Laboratory, Berkeley, California; ³Massachusetts Institute of Technology, Cambridge, Massachusetts; ⁴Pacific Northwest National Laboratory, Richland, Washington; ⁵Florida State University, Tallahassee, Florida; ⁶DOE Joint Genome Institute, Berkeley, California; ⁷National Taiwan University, Taipei, Taiwan; ⁸Desert Research Institute, Las Vegas, Nevada; ⁹University of Waterloo, London, Ontario, Canada; ¹⁰Indiana University, Bloomington, Indiana; ¹¹University of California, Berkeley, California; ¹²Howard Hughes Medical Institute, Chevy Chase, Maryland; ¹³Department of Bioengineering, University of California, Berkeley, California; and ¹⁴Princeton University, Princeton, New Jersey

Project Goals: Environmental Stress Pathway Project (ESPP) is developing computational models that describe and predict the behavior of gene regulatory networks in microbes in response to the environmental conditions found in DOE waste sites. The research takes place within the Virtual Institute for Microbial Stress and Survival (VIMSS). Based at Lawrence Berkeley National Laboratory (LBNL), VIMSS supports an integrated and multi-institutional program to understand the ability of bacteria and other microorganisms to respond to and survive external stresses. VIMSS was established in 2002 with funding from the U.S. Department of Energy Genomics:GTL Program for Rapid Deduction of Stress Response Pathways in Metal/Radionuclide Reducing Bacteria. LBNL is operated by the University of California for the U.S. Department of Energy.

A more complete picture of life on Earth, and even life *in* the Earth, has recently become possible through the application of environmental genomics. We have obtained the complete genome sequence of a new genus of the *Firmicutes*, the uncultivated sulfate reducing bacterium *Desulforudis audaxviator*, by filtering fracture water from a borehole at 2.8 km depth in a South African gold mine. The DNA was sequenced at the JGI using a combination of traditional Sanger sequencing and 454 pyrosequencing, and assembled into just one genome, indicating the planktonic community is extremely low in diversity. We analyzed the genome of *D. audaxviator* using the MicrobesOnline annotation pipeline and toolkit (<http://www.microbesonline.org>, and see MicrobesOnline abstract), which offers powerful resources for comparative genome analysis, including operon predictions and tree-based comparative genome browsing. MicrobesOnline allowed U.S. to compare the *D. audaxviator* genome with other sequenced members of the *Firmicutes* in the same clade (primarily *Pelotomaculum thermopropionicum*, *Desulfotomaculum reducens*, *Carboxydotherrmus hydrogeniformans*, and *Thermoanaerobacter tengcongensis*), as well as other known sulfate reducers (including *Archaeoglobus fulgidus* and *Desulfovibrio vulgaris*). *D. audaxviator* gives a view to the set of tools necessary for what appears to be a self-contained, independent lifestyle deep in the Earth's crust. The genome is not very streamlined, and indicates a motile, endospore forming sulfate reducer with pili that can fix its own nitrogen and carbon. *D. audaxviator* is an obligate anaerobe, and lacks obvious homologs of many of the traditional O₂ tolerance genes, consistent with the low concentration of O₂ in the fracture water and its long-term isolation from the surface. *D. audaxviator* provides a complete genome representa-

tive of the Gram-positive bacteria to further our understanding of dissimilatory sulfate reducing bacteria and archaea, and offers the full complement of genes necessary for an independent lifestyle based solely on interactions with the geochemistry of the deep subsurface.

6 ^{GT}L

Genomic Comparisons Between a Metal-Resistant Strain of *Desulfovibrio vulgaris* and the Type Strain *D. vulgaris* Hildenborough

C.B. Walker,^{1,4} D. Joyner,^{2,4} D. Chivian,^{2,4} S.S. Stolyar,^{1,4} K. Hillesland,^{1,4} J. Gabster,^{1,4} P. Dehal,^{2,4} M. Price,^{2,4} T.C. Hazen,^{2,4} **A.P. Arkin**^{2,4} (aparkin@lbl.gov), P.M. Richardson,³ D. Bruce,³ and D.A. Stahl^{1,4*}

¹University of Washington, Seattle, Washington; ²Lawrence Berkeley National Laboratory, Berkeley, California; ³DOE Joint Genome Institute, Walnut Creek, California; and ⁴Virtual Institute for Microbial Stress and Survival, <http://vimss.lbl.gov>

Project Goals: Environmental Stress Pathway Project (ESPP) is developing computational models that describe and predict the behavior of gene regulatory networks in microbes in response to the environmental conditions found in DOE waste sites. The research takes place within the Virtual Institute for Microbial Stress and Survival (VIMSS). Based at Lawrence Berkeley National Laboratory (LBNL), VIMSS supports an integrated and multi-institutional program to understand the ability of bacteria and other microorganisms to respond to and survive external stresses. VIMSS was established in 2002 with funding from the U.S. Department of Energy Genomics:GTL Program for Rapid Deduction of Stress Response Pathways in Metal/Radionuclide Reducing Bacteria. LBNL is operated by the University of California for the U.S. Department of Energy.

As part of the Virtual Institute for Microbial Stress and Survival (VIMSS), the Environmental Stress Pathway Project (ESPP) investigates the metabolic response of a sulfate-reducing bacterium, *Desulfovibrio vulgaris* Hildenborough, to contaminants found at DOE sites. Under this aegis, the ESPP Applied Environmental Core (AEC) seeks to isolate and characterize environmentally relevant sulfate-reducing bacterium from contaminated sites. Comparative analyses between these isolates and *Desulfovibrio vulgaris* Hildenborough provides an informative framework for further elucidating mechanisms of microbial adaptation to environmental stressors. To this end, a sulfate-reducing bacterium closely related to *D. vulgaris* Hildenborough was isolated from heavy-metal impacted lake sediments located in northern Illinois (Lake DePue). Initial characterization by the ESPP AEC and Functional Genomics Core (FGC) revealed differences in genome content and structure between this strain (DePue) and strain Hildenborough, despite a very high level of 16S rRNA sequence similarity (>99%). Phenotypic analyses of this strain by the AEC revealed significant differences in minimum inhibitory concentrations for a variety of compounds when compared with strain Hildenborough. Strain DePue exhibited greater tolerance towards Cr(VI) and increased sensitivity to nitrate. Small differences were observed in growth rates, although not sensitivity, for sodium between the two strains. Genome sequencing of strain DePue by the DOE Joint Genome Institute indicated that the majority of genes (approximately 90%) share a high level of similarity (>98%) to genes found in strain Hildenborough. However, the genome of strain DePue exhibits multiple genome inversions and rearrangements, as well as the presence of a several hundred novel genes not found in strain Hildenborough. Current analyses by the ESPP Computational Core (CC) verified that strain DePue lacks at least six phage regions found in strain Hildenborough, but also suggests at least two unique phage regions, one of which contains putative multi-drug efflux genes.

Further curation of the genome by the ESPP CC, as well as mutant analysis by the FGC and AEC should inform the basis for increased metal-tolerance of strain DePue and metal-resistance among *Desulfovibrio* in general.

7 ^{GTL}

Web Tools for Revealing Relationships Among Strains, Taxa, and Communities

T.G. Lilburn,¹ S.H. Harrison,^{2*} J.R. Cole,² P.R. Saxman,³ and **G.M. Garrity**² (garrity@msu.edu)

¹American Type Culture Collection, Manassas, Virginia; ²Michigan State University, East Lansing, Michigan; and ³University of Michigan, Ann Arbor, Michigan

Project Goals: The goals of this project are to develop and deploy tools that support analyses and visualizations of extremely large sequence data sets used in phylogenetic reconstructions. Current efforts are focused on validation of the self-organizing self-correcting classifier developed earlier and deployment of the tool as a web service that integrates with the RDP-II project.

Statistical approaches to understanding the species richness of prokaryotic communities in diverse environments indicate that there are thousands of yet to be cultured species (1, 4). Typically, putative members of these communities are classified and identified based on 16S rRNA genes in the extracted environmental DNA. These sequences are known as “environmental clones” in order to distinguish them from sequences from cultured organisms. The environmental clone sets are usually compared with publicly available sequences and projected as trees. In recent years the percentage of environmental clone 16S rRNA sequences in GenBank has increased from 67% to 80%. In our efforts to maintain the nomenclatural taxonomy, it has become clear that the preponderance of environmental clones is creating difficulties for researchers. We have applied our Taxomatic tool to resolving the phylogenetic taxonomy of the prokaryotes and to explore the effects of environmental clones on current classification and identification methods.

For illustrative purposes, we turn to Wagner’s recently proposed “super phylum” (5) that encompasses three recognized phyla (the *Verrucomicrobia*, the *Planctomycetes*, and the *Chlamydiae*), the *Lentisphaerae* (currently part of the *Verrucomicrobia*) and two groups that contain no cultured representatives (the OP3 candidate phylum and the *Poribacteria*). When we searched GenBank for SSU rDNA sequences affiliated with the six groups and compared them with a comparable set of sequences obtained from the RDP-II database, the differences were startling. We retrieved 3,568 SSU rDNA sequences from GenBank and 4,595 from the RDP-II database. The intersection of these two sets included only 2,160 sequences; 2,435 sequences were identified as members of this group only by the RDP-II and 1,408 were identified only by GenBank (Figure 1). Clearly, the interpretation of a community analysis would differ depending on the data set used to classify the sequences obtained from the environment.

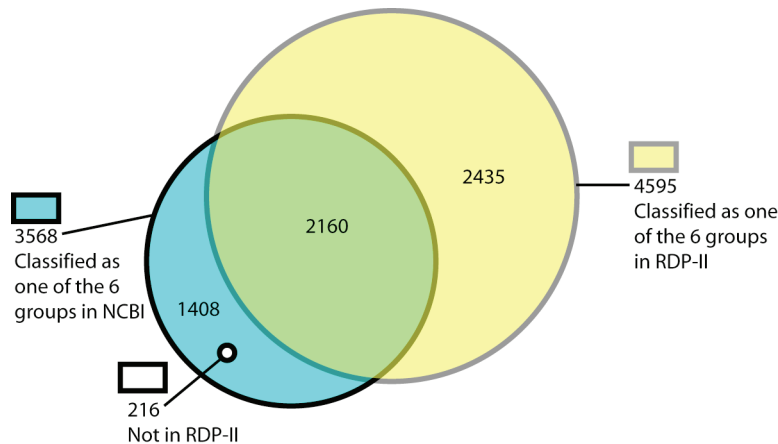


Figure 1: Showing the distribution of sequences identified as members of the six groups proposed to be members of a super phylum. Overall, 6,003 sequences were retrieved from the two databases, but only 2,160 sequences were identified by both databases as members of the six groups.

A comparison of the two classifications could help explain why there is such a marked difference in the data sets. The comparison was done using the Taxomatic, a web service built around the SOSCC algorithm (3). The algorithm produces visualizations of classifications as heat maps in R. Because biologists may not be familiar with the R or S statistical programming environments in which the algorithm runs, we have developed a web service to produce heat maps on demand. The web service has three components. The first allows users to assemble a data set and/or select a starting classification using an already-familiar interface – the RDP-II web site. The second consists of the compute engine – in essence the SOSCC algorithm running on a separate server. The third presents the results in the form of a dynamic heat map that allows users to drill down through the hierarchy and to link to external resources using digital object identifiers through NamesforLife semantic resolution services. Messaging and operations are handled by web service definition language documents in combination with the simple object access protocol. This architecture provides a robust and scalable interface to large dynamic information resources, such as the RDP-II.

Performance-tuning of the SOSCC feature set is centered around usage cases, including benchmarking, subsampling, comparing alternative hierarchies, and charting novel sequences. In the simplest case, users can upload a list of sequence identifiers to produce a publication-ready heatmap of their data. More advanced options support input of predefined or user-derived classifications. Output of intermediate steps and documentation in the SOSCC classification process are also available for user inspection. Scalability issues are also being addressed so that SOSCC web services will function as the dataset grows.

Our results show that user classifications captured by GenBank are distorting our picture of microbial diversity. Many sequences in GenBank have been placed into the taxonomic hierarchy by the submitting authors at the time of deposit, based on a BLAST-nearest neighbor approach. We have previously reported that BLAST cannot reliably determine the nearest rRNA sequence, most likely because of the high degree of similarity among all rRNA sequences (2). Moreover, since researchers apply the classification (and other annotations) from the BLAST nearest neighbor to their sequence(s), and since it is likely that the nearest neighbor is an environmental clone, each iteration of this process moves the classification another step away from any data anchored to a cultivated organism. Annotation transfer relies completely on the accuracy of prior work and therefore can

lead to misclassifications or misidentifications. This strategy also has the effect of propagating and amplifying prior errors, especially in the case of taxa for which there are few cultured representatives. In contrast, the RDP-II has implemented an on the fly Bayesian classifier to place sequences into the classification during downloading from GenBank. The classifier is trained with a carefully validated set of sequences from the nomenclatural taxonomy, so the RDP-II (and thus the Taxomatic web service) is able to provide users with more reliable and up-to-date assessments of community membership and more accurate identifications of new taxa, based on the SSU rDNA sequences.

References

1. Acinas, S. G., V. Klepac-Ceraj, D. E. Hunt, C. Pharino, I. Ceraj, D. L. Distel, and M. F. Polz. 2004. Fine-scale phylogenetic architecture of a complex bacterial community. *Nature* **430**:551-4.
2. Cole, J. R., B. Chai, R. J. Farris, Q. Wang, S. A. Kulam, D. M. McGarrell, G. M. Garrity, and J. M. Tiedje. 2005. The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Res* **33**:D294-6.
3. Garrity, G. M., and T. G. Lilburn. 2005. Self-organizing and self-correcting classifications of biological data. *Bioinformatics* **21**:2309-2314.
4. Schloss, P. D., and J. Handelsman. 2006. Toward a census of bacteria in soil. *PLoS Comput Biol* **2**:e92.
5. Wagner, M., and M. Horn. 2006. The Planctomycetes, Verrucomicrobia, Chlamydiae and sister phyla comprise a superphylum with biotechnological and medical relevance. *Curr Opin Biotechnol* **17**:241-9.

8 GTL

High Quality Microbial Finishing at JGI

Alla Lapidus^{1*} (alapidus@lbl.gov), Eugene Goltsman,¹ Steve Lowry,¹ Hui Sun,¹ Alicia Clum,¹ Stephan Trong,¹ Pat Kale,¹ Alex Copeland,¹ Patrick Chain,² Cliff Han,³ Tom Brettin,³ Jeremy Schmutz,⁴ and Paul Richardson¹

¹DOE Joint Genome Institute (JGI) Production Genomics Facility, Walnut Creek, California;

²JGI-Lawrence Livermore National Laboratory, Livermore, California; ³JGI-Los Alamos National Laboratory, Los Alamos, New Mexico; and ⁴JGI-Stanford, Stanford, California

Project Goals: To provide JGI collaborators with the highest possible quality complete microbial genomes.

The value of complete microbial genome sequence is established and appreciated by scientific community. A finished genome represents the genome assembly of high accuracy and quality (with no gaps), verified and confirmed through a number of computer and lab experiments. Several years ago JGI has established a set of high standards for the final microbial assembly and has been strictly following them thereafter.

More than 100 microbial projects have been completed since that time within the framework of the JGI's portfolio (DOE GTL program, DOE Microbial program and the Community Sequencing Program). Progress in DNA sequencing technology, design of new vectors for library construction, improvements in finishing strategies and tools, as well as the availability of a number of assemblers and advanced methods for OFR finding and genome annotation have significantly reduced the time required for genome closure. Despite this fact, complexity and speed of genome closure depends on the quality of DNA received, the whole genome shotgun libraries produced from this DNA, GC content of the genome, the size and frequency of identical or nearly identical repetitive structures,

and the amount of regions that can not be cloned or had to clone in *E. coli*. The whole genome finishing/assembly improvement pipeline will be presented showing the lab approaches and computational finishing techniques developed and implemented at JGI for finishing the large number of microbial projects in the queue. We also will present our progress in completing metagenomic projects. A number of projects for which the combination of different sequencing technologies (Sanger and 454) and finishing strategies were used will also be presented.

This work was performed under the auspices of the U.S. Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under Contract No. DE-AC02-05CH11231 and Los Alamos National Laboratory under Contract No. W-7405-ENG-36.

9 ^{GTL}

Evolution of Energy Metabolism in the *Geobacteraceae*

J.E. Butler* (jbutler@microbio.umass.edu), N.D. Young, D. Kulp, and D.R. Lovley

Department of Microbiology, University of Massachusetts, Amherst, Massachusetts

Project Goals: To determine the evolution of energy metabolism in the *Geobacter* family.

To better understand the pathways of energy metabolism in the *Geobacteraceae* family of Fe(III) reducers and electricigens, gene conservation, gene loss, and horizontal gene transfer were determined for the whole genomes of six species in the family. According to 16S rDNA phylogeny, these six species fall into two clades – the freshwater *Geobacter* clade that includes the *Geobacter* genus and *P. propionicus*, and the marine *Desulfuromonas* clade that includes *D. acetoxidans* and *P. carbinolicus*. We sought to determine the gene set shared by all members of this family, as well as to explain the presence of *Pelobacter* species, both primarily fermentative organisms, in both clades of the *Geobacteraceae*. The set of 529 proteins that were found in a single copy in each of the genomes were concatenated and used to model the phylogeny of the family. This super-tree confirmed that there are two distinct clades of *Geobacteraceae* species, and that there are *Pelobacter* species in each clade. Analysis of the genes conserved in all six genomes showed there were 713 families common to all six species. A complete TCA cycle was present in each species, although the *Pelobacters* had non-homologous isocitrate dehydrogenase and fumarase enzymes. In addition, the enzyme complexes of the inner membrane electron transport chain were generally well conserved. All species contained at least one NADH dehydrogenase, a succinate dehydrogenase, and at least one ATP synthase. Notably, only the *Pelobacters* lacked a cytochrome *bc* complex, which is predicted to move electrons from the inner membrane out to the periplasm. Surprisingly, although each genome contains at least 40 *c*-type cytochrome genes, only one of these genes was found to be conserved in all 6 members of this family, the catalytic subunit of the nitrite reductase. None of the cytochromes previously shown to be required for Fe(III) reduction or electricity generation in *G. sulfurreducens* was conserved in all the genomes. Analysis of the *Pelobacter* species indicated that they have lost a similar set of genes when compared to the non-fermenting species of the family, including: acetate transporters, all of the hydrogen-oxidizing hydrogenases, the formate dehydrogenases, and most of the *c*-type cytochromes. Analysis of gene gain in the *Pelobacter* species indicate that both gained a small cluster of dehydrogenases that allow them to use butanediol and acetoin. However, they metabolize these substrates with 2 different reaction pathways, one involving carbon-fixation and the inner membrane electron transport complexes, and the other dependent on only cytosolic proteins. These results indicate that the proto-*Geobacter* was

likely a respiring species, dependent on oxidation of carbon compounds coupled to a typical chain of electron transport complexes. The ability to ferment arose two separate times in this family.

10

Establishing Potential Chloroplast Function Through Phylogenomics

Sabeeha Merchant^{1*} (merchant@chem.ucla.edu), Steven Karpowicz,¹ Arthur Grossman,² Simon Prochnik,³ and Dan Rokhsar³

¹Department of Chemistry and Biochemistry, University of California, Los Angeles, California;

²Department of Plant Biology, The Carnegie Institution, Stanford, California; and ³DOE Joint Genome Institute, Walnut Creek, California and Center for Integrative Genomics, Department of Molecular and Cell Biology, University of California, Berkeley, California

Project Goals: The structure and metabolism of chloroplasts from green algae and land plants is strikingly well conserved, especially with respect to components of the photosynthetic apparatus and the factors involved in its assembly and maintenance. Analyses of genome sequences from various algae and plants have suggested that novel components of the chloroplast might be revealed by comparative phylogenomics. *Chlamydomonas* has served historically as a powerful model organism for the discovery of photosynthetic components and fundamental aspects of chloroplast genome expression. Therefore, we sought to use the complete set of *Chlamydomonas* protein annotations as a central data set for divining highly conserved components in the chloroplast. A list of 189 proteins conserved in organisms with green chloroplasts was generated. The goal of the project is to assess the functions of proteins on this list whose functions are not yet known. This would involve phenotypic analysis of mutants, sub-cellular locations of proteins, identification of interaction partners, and patterns of expression.

The structure and metabolism of chloroplasts from green algae and land plants is strikingly well conserved, especially with respect to components of the photosynthetic apparatus and the factors involved in its assembly and maintenance. Analyses of genome sequences from various algae and plants have suggested that novel components of the chloroplast might be revealed by comparative phylogenomics. *Chlamydomonas* has served historically as a powerful model organism for the discovery of photosynthetic components and fundamental aspects of chloroplast genome expression. Therefore, we sought to use the complete set of *Chlamydomonas* protein annotations as a central data set for divining highly conserved components in the chloroplast. The putative orthologs of *Chlamydomonas* genes from organisms with full genome sequences (*Arabidopsis*, *Physcomitrella*, diatoms, *Ostreococcus spp.*, cyanobacteria, nematode, slime mold, human, *Neurospora*, *Phytophthora*, archaea, and non-photosynthetic bacteria) were determined based on a mutual best hits approach with WU-blast (Version 2). Next, we attempted to add only close paralogs or “inparalogs” (genes that have duplicated since speciation [1]) to pairs of orthologs, while excluding ‘outparalogs’, although this is hard at such large evolutionary distances. The combination of paralogs and orthologs generated clusters of proteins that were presumably represented in the ancestor by a single gene. Having made clusters of genes from organisms across the kingdoms of life, we were able to ask for clusters that contained proteins from certain combinations of organisms, enabling U.S. to generate a series of lists of proteins conserved in different green organisms but not present in non-photosynthetic eukaryotes or prokaryotes from the protein clusters. We started with a list of 914 proteins conserved between *Arabidopsis* and *Chlamydomonas* but not present in the non-photosynthetic organisms used in the study, and then restricted the lists sequentially by inclusion of additional photosynthetic organisms, in order: *Physcomitrella*,

Ostreococcus, diatoms, *Cyanidioschizon merolae* and cyanobacteria. For *Ostreococcus*, diatoms and cyanobacteria, where genome sequences of more than one species are available, parallel lists requiring the presence of orthologs in “at least one” or “both” species were generated. A working list consisting of 189 *Chlamydomonas* proteins in 145 clusters, consisting of proteins conserved in *Arabidopsis*, *Chlamydomonas*, *Physcomitrella*, *Ostreococcus* and at least one diatom was analyzed in detail for a) known or predicted protein functions, b) predicted or experimentally-determined protein localization and c) pattern of expression of the *Arabidopsis* homolog. Most of the proteins on the list are indeed plastid localized or predicted to be so, and in most cases the pattern of expression is compatible with a function in photosynthesis or other anabolic pathways. There are no false positives on this list among the known proteins (~ 50% of the total), which include proteins unique to photosynthesis or the biogenesis of the photosynthetic apparatus, such as phosphoribulokinase, OEE proteins, Rubisco methyl transferase, HCF164, and enzymes involved in tetrapyrrole metabolism such as DVR, CHLD, GUN4, HMOX. Components unique to chloroplast metabolism were also identified among the known proteins, including proteins in the Vitamin E and Vitamin C biosynthesis pathways, and dihydropicolinate reductase in a plant specific lysine biosynthesis pathway. Plastid-specific isozymes of the pyruvate dehydrogenase complex were also selected. On this basis we conclude that the unknown components are likely to represent chloroplast-localized proteins with functions in photosynthesis or other key chloroplast metabolic pathways. Furthermore, motif analysis suggests that some of some unknown proteins may function in redox regulation, metabolite exchange, or genome maintenance/expression. Our conservative phylogenomics strategy is likely to have identified many novel proteins involved in photosynthesis, with few false hits. While the computational analysis was facilitated by the whole genome sequence of *Chlamydomonas*, the experimental accessibility of the organism means these predictions can be tested very readily.

Research sponsored by U.S. DOE Energy Biosciences, USDA NRI Plant Biochemistry, and NSF Plant Genome.

Reference

1. Sonnhammer, E.L. and E.V. Koonin, Orthology, paralogy and proposed classification for paralog subtypes. Trends Genet., 2002. 18(12): p. 619-20.

11

Beneficial Effects of Endophytic Bacteria on Biomass Production by Poplar

Safiyh Taghavi and Daniel van der Lelie* (vdlelie@bnl.gov)

Biology Department, Brookhaven National Laboratory, Upton, New York

Project Goals: The aim of this project is to understand the beneficial interaction between poplar and its endophytic bacteria. The association of endophytic bacteria with their plant hosts have been shown to have a growth-promoting effect for many different plant species. Endophytic bacteria have several mechanisms by which they can promote plant growth and health. These include the production of phytohormones or enzymes involved in growth regulator metabolism such as ethylene, 1-aminocyclopropane-1-carboxylic acid (ACC) deaminase, auxins, indoleacetic acid (IAA) or cytokinins. These mechanisms are of direct importance to the DOE mission of carbon sequestration through biomass production. In addition, endophytic bacteria can help their host plants to overcome the phytotoxic effects caused by environmental contamination, which is of direct relevance for waste management and pollution control via phytoremediation technologies.

Introduction

The association of endophytic bacteria with their plant hosts have been shown to have a growth-promoting effect for many different plant species. Endophytic bacteria have several mechanisms by which they can promote plant growth and health. These include the production of phytohormones or enzymes involved in growth regulator metabolism such as ethylene, 1-aminocyclopropane-1-carboxylic acid (ACC) deaminase, auxins, indoleacetic acid (IAA) or cytokinins. These mechanisms are of direct importance to the DOE mission of carbon sequestration through biomass production. In addition, endophytic bacteria can help their host plants to overcome the phytotoxic effects caused by environmental contamination, which is of direct relevance for waste management and pollution control via phytoremediation technologies.

Scientific approach

Recent research by our group has illustrated the potential of endophytic bacteria to increase the net primary biomass production of the host *Populus* tree. The goal of our research is to identify specific strains of endophytic bacteria that improve the growth and carbon sequestration potential of *Populus*. We therefore conducted a high-throughput screen of approximately 100 bacterial endophyte strains to identify those strains with the greatest impact on *Populus* net primary productivity. As an example, the results of inoculation of poplar with 8 different endophytic strains are presented. The most significant stimulation in growth was observed with *Enterobacter* sp. 638*, followed by *Burkholderia cepacia* L.S.2.4 and *Stenotrophomonas maltophilia* R551-3*. Some strains had no significant effect on biomass production as compared to non inoculated control plants. This was the case for *Enterobacter* sp. R558-1, *Pseudomonas putida* W619*, *Serratia proteamaculans* 568* and plants inoculated with the soil bacterium *Ralstonia metallidurans* CH34 (control). Interestingly, plants inoculated with *Methylobacterium populi* BJ001* showed a strong reduction in growth, despite the fact that this strain was isolated as an endophyte from poplar tissue cultures.

To better understand the interactions between poplar and its endophytic bacteria we initiated in collaboration with DOE's JGI the full genome sequencing of 5 endophytic strains (marked by *). A first analysis of the draft genome sequences resulted in the identification of several functions that would allow the endophytic bacteria to interact with the development of their poplar host. Several strains contained a copy of a 1-aminocyclopropane-1-carboxylic acid (ACC) deaminase gene, seemed to be able to produce indoleacetic acid (IAA), or metabolize phenyl acetic acid. In addition, *Pseudomonas putida* W619 was shown to contain the uptake carrier for 4-amino butyrate, another important plant hormone. Genome comparison between the endophytes and closely related non-endophytic strains from the same species should provide U.S. with valuable insights about the essential functions for successful endophytic colonization by these bacteria of their poplar host.

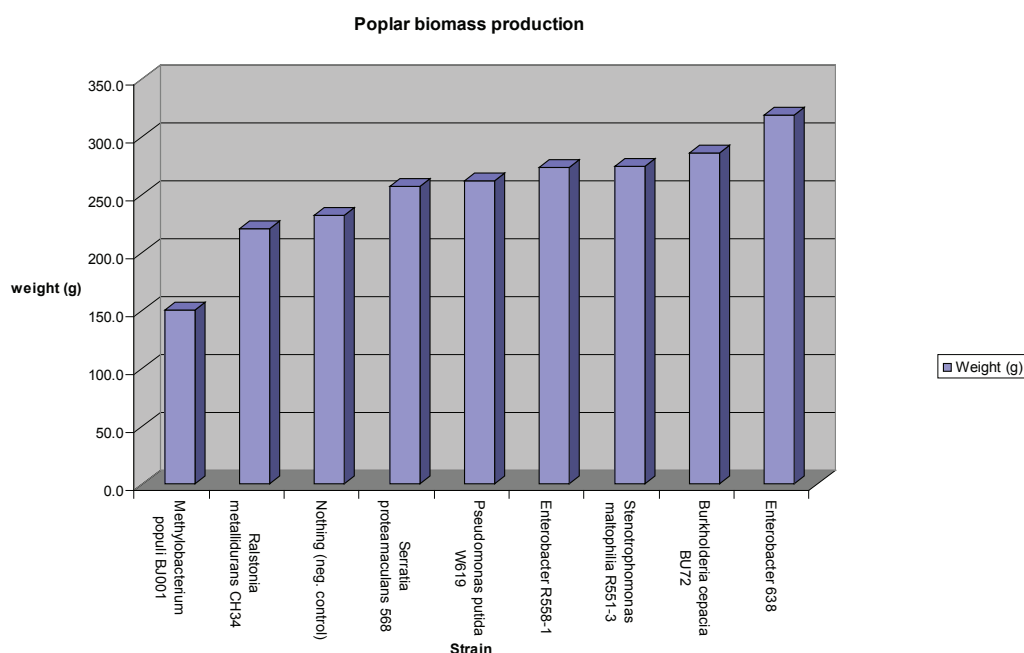


Figure 1: beneficial effects of endophytic colonization on the biomass production of their poplar host. Results are the average of 7 poplar trees per condition.

Future work

We will test the hypothesis that the interacting genomes of bacterial endophytes and *Populus* not only increase *Populus* biomass production, but also increase partitioning of resources into long-lived, i.e. recalcitrant forms of carbon. In parallel, endophyte properties that are hypothesized to be important for colonization and plant growth promotion will be investigated, including capacity for production of extracellular hydrolytic enzymes, nitrogen-fixing enzymes, and low molecular weight compounds such as phytohormones. Also, the dynamics of endophytic colonization will be explored using GFP-expressing endophytic strains. With these data in hand we will embark upon a systems biology approach to better understand the interaction between endophytic bacteria and their *Populus* host. The resulting comprehensive view of the endophyte-*Populus* interacting genomes has the potential to be used in developing recommendation in use of endophyte inoculant to increase carbon sequestration in *Populus* plantations.

Acknowledgement

This work was supported by the U.S. Department of Energy, Office of Science, BER, project number KP1102010. This work was also funded under Laboratory Directed Research and Development project number LDRD05-063. Sequencing of the endophytic genomes is been carried out at the Joint Genome Institute (JGI) under the auspices of the U.S. Department of Energy's Office of Science, Biological and Environmental Research Program.

Section 2

Microbial Community Sequencing and Analysis

12 ^{GTL}

Structure and Dynamics of Natural Low-Diversity Microbial Communities

Jillian F. Banfield^{1,5*} (jill@eps.berkeley.edu), Vincent Denef,¹ Nathan VerBerkmoes,² Paul Wilmes,¹ Gene Tyson,¹ John Eppley,¹ Genevieve DiBartolo,¹ Daniela Goltsman,¹ Anders Andersson,¹ Chris Belnap,¹ Brett J. Baker,¹ Linda Kalnejais,¹ A. Pepper Yelton,¹ D. Kirk Nordstrom,³ Eric E. Allen,¹ Rachel Whitaker,¹ Sheri Simmons,¹ Manesh Shah,² Michael Thelen,⁴ Gary Andersen,⁵ and Robert Hettich²

¹University of California, Berkeley, California; ²Oak Ridge National Laboratory, Oak Ridge, Tennessee; ³U.S. Geological Survey, Boulder, Colorado; ⁴Lawrence Livermore National Laboratory, Livermore, California; and ⁵Lawrence Berkeley National Laboratory, Berkeley, California

Project Goals: The interdisciplinary research in our GTL project is motivated by the desire to understand how microbial communities assemble, the interplay amongst community members, and the mechanisms of microbial adaptation and evolution. Chemoautotrophic acid mine drainage (AMD) microbial biofilms have proven particularly tractable for these analyses due to their comparatively low species richness. The goal of this subproject is to develop experimental and computational approaches for the comprehensive characterization of the proteome of the AMD system to investigate the nature of the gene expression and conservation amongst the various microbial members of this consortium. Proteomic information will be integrated with genomic and biochemical datasets to help elucidate the structure and activity of microbial communities in their natural environmental context.

The interdisciplinary research in our GTL project is motivated by the desire to understand how microbial communities assemble, the interplay amongst community members, and the mechanisms of microbial adaptation and evolution. Our approach is to use cultivation-independent methods to study the structure and activity of microorganisms in their natural environmental context. Chemoautotrophic acid mine drainage (AMD) microbial biofilms have proven particularly tractable for these analyses due to their comparatively low species richness. Recent advances have relied upon the development and application of community genomic and proteomic methods, as well as new contextual information provided by geochemical analyses and ultrastructural characterization.

Genomic studies reveal that AMD biofilms are typically dominated by near clonal bacterial populations. Spatial and temporal analyses indicate essentially a single species of *Leptospirillum* group III, the only organism capable of nitrogen fixation, across a diversity of environment types. In contrast, different biofilms are dominated by one of a set of *Leptospirillum* group II genomes formed by homologous recombination between two end member genome types. Analyses of biofilm growth stages suggest selection for a UBA *Leptospirillum* group II type during initial colonization, introduction of archaea and fungi in intermediate succession stages, and dominance by the 5-way CG *Leptospirillum* group II genome type in mature biofilms.

Within bacterial populations, the predominant form of genomic heterogeneity is in gene content. These differences are particularly prevalent in regions impacted by phage and plasmid integration. In some cases, these regions encode key functions, including cytochrome variants and quorum sensing genes. Our findings further indicate that subsets of these genes are under selection. Variation in metabolic potential resulting from gain or loss of phage-related genes is inferred to be particularly important in diversification of otherwise near clonal *Leptospirillum* group III populations. Predominance of a single genome type for each bacterial species may reflect strong clonal expansion events, particularly early in the early colonization of the air-solution interface.

Archaea are numerically less abundant in biofilm communities. Archaeal populations typically have complex genome pools consisting of combinatorial variants, and also exhibit heterogeneity in gene content concentrated in phage insertion regions. The high degree of sequence-level variation is maintained by rapid rates of homologous recombination, possibly ensuring a continuum of adaptation potential. This may be important, given that they appear in successional stages where microenvironmental heterogeneity is likely the result of established biofilm architecture.

In addition to assembly of microbial consortia in response to physical and chemical conditions and biofilm growth stage, recent evidence suggest that viral predation is an important selective force that shapes microbial consortia and drives their evolution. We find evidence for dramatic crashes in biofilm communities, possibly caused by phage blooms, accompanied by a switch in the dominant bacterial strain type. Periodic decimation of the dominant populations is not unexpected (*kill the king*), given ongoing microbial evolution to outwit phage counterbalanced by phage evolution to evade host defenses.

Microbial genomes may provide insight into the mechanisms of phage defense and record information about the recent history of phage predation. All of the microorganisms in the AMD biofilms contain at least one region of short tandem repeats separated by similar length spacer sequences (CRISPR), accompanied by a set of CRISPR-associated proteins; previous studies have suggested that this comprises a microbial immune system. We find very high levels of population heterogeneity in the spacers between tandem repeats, consistent with rapid diversification of the inventory of possible RNAi-like molecules available to silence foreign DNA. A significant subset of the spacers shows sequence similarity to transposase genes, prophage genes, and unassembled sequencing reads (possible derived from phage). In combination with other data, results support a role of spacers in phage defense. Extremely rapid CRISPR evolution is expected if the region is responding to a rapidly changing selection pressure associated with phage predation. Bacterial CRISPR-associated proteins are some of the most abundant proteins in the biofilms, reinforcing the importance of these large genomic loci to organism survival. Ongoing parallel studies of phage communities, in combination with genomic and proteomic analysis of bacteria and archaea, will be vital for development of a more detailed understanding of microbial community dynamics.

This research sponsored by the U.S. DOE-BER, Genomics:Genomics:GTL Program.

13 ^{GTL}

A Novel Binning Approach and Its Application to a Metagenome From a Multiple Extreme Environment

N. Maltsev^{1*} (maltsev@mcs.anl.gov), M. Syed,¹ A. Rodriguez,¹ B. Gopalan,² and **F. Brockman**²

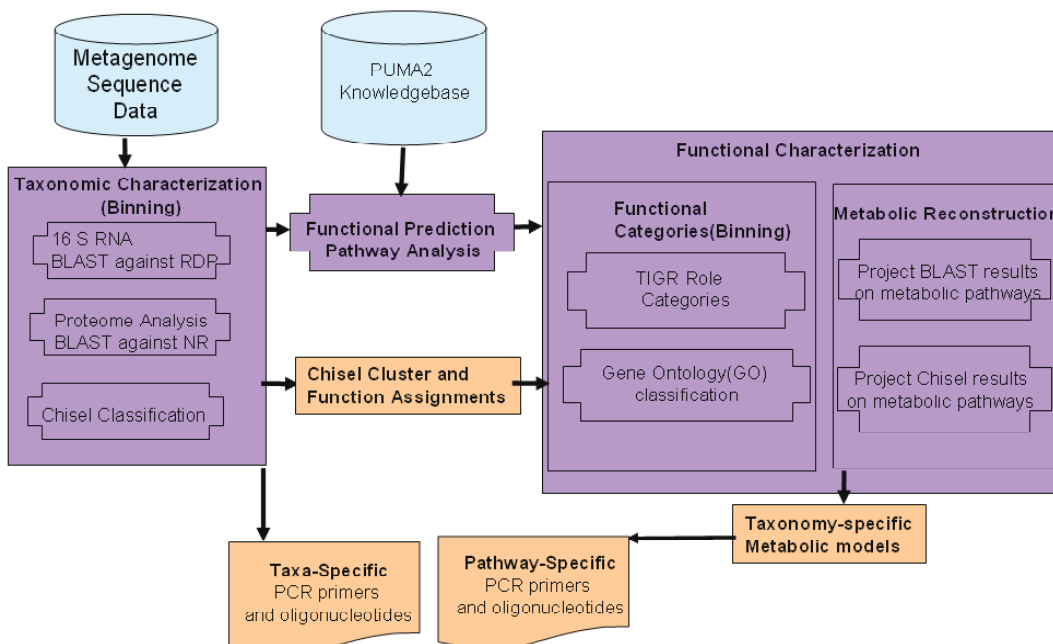
¹Argonne National Laboratory, Argonne, Illinois and ²Pacific Northwest National Laboratory, Richland, Washington

Project Goals: A metagenome of the living microbial community present in low biomass subsurface sediments beneath a leaking high-level radioactive waste tank at the DOE Hanford Site was analyzed using advanced bioinformatics methods. Besides very high levels of radiation this microbial community was also subjected to high temperatures and dessication, extremely high concentrations of chromium and nitrate, and alkaline conditions. The great majority of microbes in most natural environments will be represented in metagenome sequence by a limited number of genes which do not assemble, due to substantial community diversity and limited sequencing depth. Therefore, a critical need is more accurate and informative binning of genes into taxonomic groups, to enable improved reconstruction of the metabolic and physiological processes operating in the community. We have developed a new approach for binning metagenome sequences, and applied it to a low-biomass microbial community exposed for several decades to multiple extreme conditions. We have discovered that many of the proteins found in this metagenome show homology to those found in extremophilic microbes, indicating that the community has undergone systems-level changes.

A metagenome of the living microbial community present in low biomass subsurface sediments beneath a leaking high-level radioactive waste tank at the DOE Hanford Site was analyzed using advanced bioinformatics methods. Besides very high levels of radiation this microbial community was also subjected to high temperatures and dessication, extremely high concentrations of chromium and nitrate, and alkaline conditions. The great majority of microbes in most natural environments will be represented in metagenome sequence by a limited number of genes which do not assemble, due to substantial community diversity and limited sequencing depth. Therefore, a critical need is more accurate and informative binning of genes into taxonomic groups, to enable improved reconstruction of the metabolic and physiological processes operating in the community. We have developed a new approach for binning metagenome sequences, and applied it to a low-biomass microbial community exposed for several decades to multiple extreme conditions. We have discovered that many of the proteins found in this metagenome show homology to those found in extremophilic microbes, indicating that the community has undergone systems-level changes.

Hanford site metagenome. Biomass in the sediments is present in very low quantities (~10,000 cells per gram), about one-millionth the levels routinely found in soils. In order to exclude the large fraction of dead cells and to provide adequate biomass for library construction, many culture enrichments were pooled. Low-coverage shotgun sequencing was performed by the DOE Production Genomics Facility. Metagenome sequence obtained from pooled enrichments from the extreme contamination and low contamination zones were separately analyzed using bioinformatics approaches to achieve the following: 1. Determine the taxonomy of organisms selected for by the extremophilic conditions 2. Reconstruct major physiological properties of the microbial community from available genomic data and 3. Identify the sequence features associated with extremophilic phenotypes.

The Schema Of Hanford Site Metagenome analysis



1. Taxonomic Profiling (Binning) was done using: a) Phylogenetic analysis of the 16S rDNA b) Taxonomic analysis of the BlastX results and c) Identification of taxonomic variations of enzymes using the Chisel system being developed by our group. This unique step allows increasing resolution and reliability of predictions of enzymatic capabilities characteristic for particular taxonomic groups in the samples.

2. Reconstruction of Physiological Profiles for taxonomic groups of the microbial community. The gene function predictions were based on the results of analysis of metagenomes using Chisel [<http://compbio.mcs.anl.gov/CHISEL>] for identification of taxonomy and phenotype-specific variations of enzymatic sequences, as well as by traditional bioinformatics tools (e.g. BlastX, InterPro, Blocks). The predicted functions attributed to particular organisms or taxonomic groups were projected onto the library of metabolic pathways from the EMP and KEGG databases. Reconstructions of signal transduction and transmembrane transport systems were also performed.

Results

Sixty-three 16S rDNA and 13,388 protein sequences from the extreme and low contamination zones were analyzed. According to the analysis of the 16S rDNA the extreme contamination zone was dominated by *Deinococcus-Thermus*, *Actinobacteria* and *Firmicutes*. The *Actinobacteria* were represented by *Micrococccineae*, *Propionibacterineae*, *Corynebacterineae* and *Frankineae* suborders. Taxonomic analysis of Chisel results and BlastX hits has confirmed that this population was dominated by the *Deinococcus-Thermus* phylum (~60% of ORFs) and *Actinobacteria* (~37% of ORFs). A significant number of homologs to sequences from extremophilic organisms were identified. The low contamination zone was dominated by *Proteobacteria* and *Actinobacteria*; 6 of 9 16S rDNA sequences were attributed to *Actinobacteria*. Analysis of protein sequences using Blast and Chisel has attributed ~ 55% of translated ORFs to *Proteobacteria* phylum and 40% to *Actinobacteria*.

Metabolic Reconstructions from sequence data were done using the gene function predictions based on Chisel results and conventional bioinformatics tools. Due to low coverage of genomes in the samples, in the majority of cases reconstruction of physiological profiles for individual species was impossible. Therefore hierarchical reconstructions for higher taxonomic groups (e.g. genus, order) were performed. The predicted enzymes in the most abundant groups (*Actinomycetales* and *Bacillus*) corresponded to the core metabolic pathways. An interesting finding was the identification of enzymes in the streptomycin biosynthetic pathway in *Actinobacteria*. Streptomycin is known to induce streptomycin-dependent error-prone protein biosynthesis, that may be advantageous for microorganisms residing in extreme conditions.

Evolutionary Analysis of genomes of extremophilic organisms. The Chisel analysis of the extreme contamination zone metagenome identified 543 enzymatic sequences corresponding to 263 distinct enzymatic functions. The predominant taxonomic groups of organisms identified in the sample were *Actinomycetales* (28%, corresponding to 152 Chisel predictions) and *Bacillus* (22%, corresponding to 122 Chisel predictions). Other predicted groups included a number of hits from extremophilic organisms: *Deinococcus*, *Euryarchaeota*, and *Symbiobacterium thermophilum*. These results match the results predicted by 16S rDNA analysis of this data. Chisel allows for further investigation of this metagenome by supporting the design of taxonomy-specific oligonucleotides for messenger RNA-targeted fluorescence in situ hybridization (FISH) studies. These degenerative oligonucleotides are based on the alignments of sequences corresponding to taxonomy-specific Chisel clusters.

The taxonomic profile of the microbial community identified in the extreme contamination zone shows a surprising similarity to the community in untreated and low radiation (0.5 MRad) treated soils of the Atacama desert (Rainey et al., 2004). Both populations were dominated by *Actinobacteria*, *Deinococcus* and *Firmicutes*. This observation leads to the suggestion that microbial populations residing in extremophilic natural environments are pre-conditioned for adapting to and surviving new, human-caused extremophilic conditions. To test this hypothesis we further analyzed the taxonomy-specific variations of enzymes identified by Chisel in our metagenomes, using high-resolution bioinformatics tools developed by our group (e.g. Dragonfly and Phyloblocks) for evolutionary analysis of protein sequences. Our analysis shows divergent evolution of enzymatic functions that lead to the emergence of Actinobacterial and Deinococcal variations of some essential enzymes of glycolysis, nucleotide biosynthesis, DNA repair systems, and others. Many of these enzymes also show subsequent convergent evolutionary changes characteristic for extremophilic microbes from different taxonomic groups. We conclude that in the course of adaptation to the conditions in the Hanford sediments, the community has undergone systems-level changes spanning multiple biological functions.

14 ^{GTL}

Insights into Stress Ecology and Evolution of Microbial Communities from Uranium-Contaminated Groundwater Revealed by Metagenomics Analyses

Christopher L. Hemme,^{1,6,8*} Ye Deng,¹ Terry Gentry,⁶ Liyou Wu,¹ Matthew W. Fields,^{2,8} David Bruce,³ Chris Detter,³ Kerrie Barry,³ David Watson,⁶ Paul Richardson,³ James Bristow,³ Terry C. Hazen,^{4,8} James Tiedje,⁵ Eddy Rubin,³ **Adam P. Arkin**^{7,8} (aparkin@lbl.gov), and Jizhong Zhou^{1,8}

¹Institute for Environmental Genomics, Department of Botany and Microbiology, University of Oklahoma, Norman, Oklahoma; ²Department of Microbiology, Miami University, Oxford, Ohio; ³DOE Joint Genome Institute, Walnut Creek, California; ⁴Earth Science Division, Lawrence Berkeley National Laboratory, Berkeley, California; ⁵Center for Microbial Ecology, Michigan State University, East Lansing, Michigan; ⁶Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee; ⁷Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California; and ⁸Virtual Institute for Microbial Stress and Survival, <http://vimss.lbl.gov>

Project Goals: Environmental Stress Pathway Project (ESPP) is developing computational models that describe and predict the behavior of gene regulatory networks in microbes in response to the environmental conditions found in DOE waste sites. The research takes place within the Virtual Institute for Microbial Stress and Survival (VIMSS). Based at Lawrence Berkeley National Laboratory (LBNL), VIMSS supports an integrated and multi-institutional program to understand the ability of bacteria and other microorganisms to respond to and survive external stresses. VIMSS was established in 2002 with funding from the U.S. Department of Energy Genomics:GTL Program for Rapid Deduction of Stress Response Pathways in Metal/Radionuclide Reducing Bacteria. LBNL is operated by the University of California for the U.S. Department of Energy.

One of the central goals of ESPP is to understand the responses of microbial communities to various stresses within the context of field applications. Towards this goal, we are sequencing groundwater microbial communities with manageable diversity and complexity (~10-400 phylotypes) at the U.S. Department of Energy's Environmental Remediation Science Program (ERSP)-Field Research Center (FRC), Oak Ridge, TN. The microbial community has been sequenced from a groundwater sample (FW106) contaminated with very high levels of nitrate, uranium and other heavy metals and pH ~3.7. Consistent with trends expected in stressed ecosystems, the metagenome reveals a community of low species and strain diversity dominated by a single *Frateriella*-like γ -proteobacteria with other γ - and β -proteobacteria present at low proportions. Metabolic reconstruction reveals specific adaptations to the geochemical conditions of FW106 including genes encoding metal resistance (*czcABC*, *czcD*, *cadA*, *merA*, *arsB*), denitrification, and solvent resistance (1,2-dichloroethene, acetone, butanol). In addition to the presence of these specific genes, certain resistance genes also appear to be overrepresented in the metagenome including genes from nitrate/nitrite transport (*narK*) and metal translocation (*czcABC*, *czcD*, *cadA*), likely due to a combination of gene duplication and lateral gene transfer. A screen for positive selection shows most of these genes to be under strong negative selection, suggesting that in the short term at least, the overabundance of these transporters provide a positive fitness benefit to the cell by increasing the rate of ion transport. SNP analysis revealed a low level of polymorphism with the overwhelming majority of SNP representing unique changes within the assembled reads, suggesting that the strains in the sample are largely clonal. A model is presented for the evolution of microbial communities under high-stress conditions. To understand the metabolic diversity of the groundwater microbial community, the microbial community (~400 phylotypes) from the background well at the FRC is also currently under sequencing.

15 ^{GTL}**Changes in Microbial Community Structure During Biostimulation for Uranium Reduction at Different Levels of Resolution**

C. Hwang,^{1,8*} W.-M. Wu,² T.J. Gentry,³ J. Carley,⁴ S.L. Carroll,⁴ D. Watson,⁴ P.M. Jardine,⁴ J. Zhou,^{5,8} T.C. Hazen,^{6,8} E.L. Brodie,^{6,8} Y.M. Piceno,⁶ G.L. Andersen,⁶ E.X. Perez,⁷ A. Masol,⁷ C.S. Criddle,² and M.W. Fields^{1,8}

¹Department of Microbiology, Miami University, Oxford, Ohio; ²Department of Civil and Environmental Engineering, Stanford University, Stanford, California; ³Department of Crop and Soil Sciences, Texas A & M University, College Station, Texas; ⁴Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee; ⁵Institute for Environmental Genomics, University of Oklahoma, Norman, Oklahoma; ⁶Center for Environmental Biotechnology, Lawrence Berkeley National Laboratory, Berkeley, California; ⁷Department of Biology, University of Puerto Rico, Mayaguez, Puerto Rico; and ⁸Virtual Institute for Microbial Stress and Survival (<http://vimss.lbl.gov/>)

Project Goals: Environmental Stress Pathway Project (ESPP) is developing computational models that describe and predict the behavior of gene regulatory networks in microbes in response to the environmental conditions found in DOE waste sites. The research takes place within the Virtual Institute for Microbial Stress and Survival (VIMSS). Based at Lawrence Berkeley National Laboratory (LBNL), VIMSS supports an integrated and multi-institutional program to understand the ability of bacteria and other microorganisms to respond to and survive external stresses. VIMSS was established in 2002 with funding from the U.S. Department of Energy Genomics:GTL Program for Rapid Deduction of Stress Response Pathways in Metal/Radionuclide Reducing Bacteria. LBNL is operated by the University of California for the U.S. Department of Energy.

Former radionuclide waste ponds at the ERSP-Field Research Center in Oak Ridge, TN pose several challenges for uranium bioremediation. The site is marked by acidic conditions, high concentrations of nitrate, chlorinated solvents, and heavy metals. Bacterial sulfate reduction can be an important process for the bio-reduction of *in situ* heavy metals, but little is known how potential stressors can impact bio-reduction activities at the cellular, population, and community levels. A goal of VIMSS is to characterize ecosystem responses at coordinated levels of resolution in order to predict cellular responses at DOE waste sites. Through VIMSS efforts, population and community level responses can be correlated with cellular responses from individual stress experiments, and this work allows for a more complete understanding of the system. The current work uses a series of re-circulating wells serve to create a subsurface bioreactor to stimulate microbial growth for *in situ* U(VI) immobilization (Wu et al. ES&T 40:3986-3995). Well FW-104 is the injection well for the electron donor (i.e., ethanol); well FW-026 is the extraction well for the recirculation loop; well FW-101 and FW-102 are the inner zones of biostimulation; and FW-024 and FW-103 are upstream and downstream wells, respectively, which are the outer protective zones. Microbial community composition and structure of the groundwater from the wells were analyzed via clonal libraries of partial SSU rRNA gene sequences, a phylogenetic chip array (Bacteria and Archaea), and a functional gene chip array over time. LIBSHUFF analysis for the clonal libraries of the re-circulating wells showed that over each phase of manipulation for uranium immobilization, the bacterial communities of the inner zones of biostimulation were more similar to each other than those of the outer protective zones. The outer protective zones were more similar to the injection well. LIBSHUFF analyses for the clonal libraries from FW-104 (injection), FW-101 and FW-102 (biostimulation) showed that bacterial communities of the three wells were initially similar but developed changes through time. FW-101

and FW-102 bacterial communities developed changes in parallel, while those of FW-104 showed gradual changes. Diversity indices showed that bacterial diversity tended to increase during the initial phase of uranium bioreduction and decreased toward the end of uranium bioreduction (i.e., low U(VI) levels). In addition, when electron donor was added to the subsurface, community diversity increased with a subsequent decline in U(VI) levels. However, when levels of potential electron acceptors decreased, community diversity also decreased. As uranium levels declined, increasing *Desulfovibrio*- and *Geobacter*-like sequences were detected from the clonal libraries; moreover, *Desulfovibrio*-like sequences predominated over time. The results were further confirmed via RT-PCR, and RT-PCR results correlated with OTU and PhyloChip distributions for *Desulfovibrio*. PhyloChip analyses also demonstrated the presence and dynamics of both acetoclastic and hydrogenotrophic methanogens. The microbial community dynamics from one of the 4 frequently sampled monitoring wells (FW 102-3) was intensively analyzed with a functional gene array containing 27,000 probes covering 10,000 genes and >100 gene categories. The microarray data indicated that during the uranium reduction period, both FeRB and SRB populations reached their highest levels at Day 212, followed by a gradual decrease over 500 days. The uranium concentrations in the groundwater were significantly correlated with total abundance of c-type cytochrome genes ($r=0.73$, $p<0.05$) from *Geobacter*-type FeRB and *Desulfovibrio*-type SRB, and with the total abundance of *dsrAB* (dissimilatory sulfite reductase) genes ($r=0.88$, $p<0.05$). Mantel test of microarray data and chemical data also indicated that there was significant correlation between the differences of uranium concentrations and those of total c-cytochrome gene abundance ($r=0.75$, $p<0.001$) or *dsrAB* gene abundance ($r=0.72$, $p<0.01$). The changes of more than a dozen individual c-type cytochrome genes from *Geobacter sulfurreducens* and *Desulfovibrio desulfuricans* showed significant correlations to the changes of uranium concentrations among different time points. Also the changes of more than 10 *dsrAB*-containing populations, including both cultured (e.g. *Desulfovibrio* spp., *Desulfotomaculum*, and *Thermosedulfovibrio*) and non-cultured SRB were significantly related to the changes in uranium concentrations. These results suggested the importance of these functions for *in situ* uranium reduction. Interestingly, the changes of several *dsrAB* sequences previously recovered from this site (e.g., FW003269B, FW300181B) showed significant correlations to the changes in uranium levels. In conclusion, the microbial community composition and structure changed upon stimulating for uranium bioreduction conditions, and that sequences representative of the sulfate-reducers *Desulfovibrio* spp. and metal-reducers *Geobacter* spp. were detected in wells that displayed a decline in U(VI).

16 ^{GTL}**VIMSS Applied Environmental Microbiology Core Research on Stress Response Pathways in Metal-Reducers**

Terry C. Hazen,^{1,9*} Carl Abulencia,^{3,9} Gary Anderson,^{1,9} Sharon Borglin,^{1,9} Eoin Brodie,^{1,9} Steve van Dien,⁷ Matthew Fields,^{6,9} Jil Geller,^{1,9} Hoi-Ying Holman,¹ Richard Phan,^{1,9} Eleanor Wozel,^{1,9} Janet Jacobsen,^{1,9} Dominique Joyner,^{1,9} Romy Chakraborty,^{1,9} Martin Keller,^{2,9} Aindrila Mukhopadhyay,^{1,9} David Stahl,^{5,9} Sergey Stolyar,^{5,9} Judy Wall,^{4,9} Huei-che Yen,^{4,9} Grant Zane,^{4,9} Jizhong Zhou,^{8,9} E. Hendrickson,^{5,9} T. Lie,^{5,9} J. Leigh,^{5,9} and Chris Walker^{5,9}

¹Lawrence Berkeley National Laboratory, Berkeley, California; ²Oak Ridge National Laboratory, Oak Ridge, Tennessee; ³Diversa, Inc., San Diego, California; ⁴University of Missouri, Columbia, Missouri; ⁵University of Washington, Seattle, Washington; ⁶Miami University, Oxford, Ohio; ⁷Genomatica, San Diego, California; ⁸University of Oklahoma, Norman, Oklahoma; and ⁹Virtual Institute for Microbial Stress and Survival, <http://vimss.lbl.gov>

Project Goals: Environmental Stress Pathway Project (ESPP) is developing computational models that describe and predict the behavior of gene regulatory networks in microbes in response to the environmental conditions found in DOE waste sites. The research takes place within the Virtual Institute for Microbial Stress and Survival (VIMSS). Based at Lawrence Berkeley National Laboratory (LBNL), VIMSS supports an integrated and multi-institutional program to understand the ability of bacteria and other microorganisms to respond to and survive external stresses. VIMSS was established in 2002 with funding from the U.S. Department of Energy Genomics: GTL Program for Rapid Deduction of Stress Response Pathways in Metal/Radionuclide Reducing Bacteria. LBNL is operated by the University of California for the U.S. Department of Energy.

Field Studies

Environmental Characterizations. Clonal libraries for the SSU rRNA gene are a commonly used tool for the characterization of bacterial communities, and confidence intervals were predicted for accuracy of sequence determination from SSU rRNA libraries. The data and the model results suggested that similarity values below 0.995 are likely derived from dissimilar sequences at a confidence level of 0.95, and not sequencing errors. The results confirmed that screening by direct sequence determination could be reliably used to differentiate at the species level (Fields et al., 2006). Clonal libraries were then used to characterize changes in community structure along a contaminant plume (Oak Ridge, TN) in terms of phylogenetic, functional, and geochemical changes. Such studies are essential to understand how a microbial ecosystem responds to perturbations. Our results indicated that different gene sequences estimated different relationships between populations within the microbial communities. However, functional groups that respond differently under a particular perturbation should have different patterns of diversity along the contaminant gradient in relation to growth and competitive displacement, and the data supported this hypothesis (Fields et al., 2006b). An additional study characterized the community changes in a fluidized bed reactor for the treatment of uranium-contaminated groundwater. Changes in community structure and composition were correlated to operating conditions, and relationships between diversity and stability were discussed (Hwang et al., 2006). Our current work has been the identification of predominant populations in the uranium/nitrate-contaminated subsurface during bio-stimulation for heavy metal reduction. The data, thus far, indicated that when electron donor was added to the subsurface, community diversity increased with a subsequent decline in U(VI) levels. However, when levels of potential electron acceptors decreased, community diversity also decreased. As uranium levels declined, increasing *Desulfovibrio* and *Geo-*

bacter-like sequences were detected from the clonal libraries; moreover, *Desulfovibrio*-like sequences predominated over time (Hwang et al., 2006b). Previous research specifically points toward SRB as environmentally relevant experimental systems for the study of heavy metal and radionuclide reduction, and our recent data has detected *Desulfovibrio* sequences at the FRC and Hanford. To effectively immobilize heavy metals and radionuclides, it is important to understand the cellular responses to adverse factors observed at contaminated subsurface environments, such as mixed contaminants and the changing ratios of electron donors and acceptors. In a recent study, we focused on stasis-induced genes and gene networks by looking at transition of *D. vulgaris* from exponential- to stationary-phase during electron donor depletion. Our results demonstrated that *D. vulgaris* cells altered gene expression profiles in response to carbon and energy depletion, and that gene expression during stationary-phase was not static. In particular, genes related to phage, carbohydrate flux, outer envelop, and iron homeostasis played a major role in the cellular response to nutrient deprivation under the tested growth conditions.

^{13}C -labelled lactate was injected in August 2004 at the Hanford 100H site to biostimulate chromium reduction. After more than 1 year, chromium was still at non-detect in the stimulated wells. 16s phylochip analyses showed a dramatic increase in diversity at the stimulated wells, including iron reducers (*Geobacter*) and sulfate reducers (*Desulfovibrio*). Sequentially competing terminal electron acceptors were depleted: oxygen, nitrate, iron(III), and sulfate. Methane however was never detected, though ^{13}C was detected in the dissolved inorganic carbon and in the signature lipids (PLFA) of iron reducers and sulfate reducers. Sulfate reduction was still active after two years in the deepest parts of the aquifer, and iron(II) still dominated suggesting an active Cr(IV) reducing environment. *Desulfovibrio* strains have been isolated and are currently being sequenced. Stress responses in these strains will be compared to the pipeline studies on DvH already completed.

Biopanning/Clone libraries. This year we did further optimization of the MDA approach to isolate and amplify DNA from samples with extreme low biomass. We did a workup on the Hanford samples to construct environmental libraries for sequencing and screening. We also evaluated three different methods to remove rRNA and tRNA from samples. The first method utilizes biotin-modified oligos complementary to conserved regions in 16S & 23S rRNA and subtractive hybridization with streptavidin-coated magnetic beads. The second uses a commercially available exonuclease that specifically digests rRNAs bearing a 5' monophosphate group. The third method uses two rounds of reverse transcription, where rRNAs are first reverse transcribed with multiple universal primers for 16S & 23S RNAs, subsequently the RNA/DNA hybrids and cDNA are removed by sequential digestion with RNaseH and DNaseI, and the enriched mRNAs are then reverse transcribed using random primers. We evaluated these three methods by comparing disappearance of the 16S and 23S bands via electrophoresis, and their effect on mRNA quality and quantity by analysis of transcription levels of control (total RNA) vs. enriched mRNA as measured whole genome microarray. Enriched mRNAs from the first two methods generated more genes with altered transcript levels compared to untreated total RNA, with 19 genes (0.5%) for the exonuclease method & 74 genes (2%) for subtractive hybridization exhibiting significant differences ($P < 0.05$).

Genome Sequence. The genome sequence for *Desulfovibrio vulgaris* DePue strain DP4 has been closed and is now being annotated in collaboration with other ESPP investigators (see other abstract).

Dual culture systems. We achieved steady state growth of a syntrophic association between *Desulfovibrio vulgaris* and *Methanococcus maripaludis* in chemostats equipped with on-line monitoring of volatile metabolites (hydrogen, methane, CO_2). In association with other ESPP investigators, transcriptional analyses of replicated chemostat-grown cocultures and monocultures were completed (see other abstracts). Characterization of the corresponding proteomes is now in progress. In collaboration with the Wall laboratory, mutants in genes implicated in syntrophic growth were examined

for growth in monoculture and in coculture (see other abstracts). These experiments strongly suggest that the Coo Hydrogenase plays a central role in energy conservation during syntrophic growth, possibly functioning as a proton translocating hydrogenase. A second candidate energy converting hydrogenase, Ech, was demonstrated to play a minor role in syntrophic growth, but has been implicated in the production of reduced ferredoxin required for synthesis of pyruvate when growing on hydrogen and acetate.

Stress Experiments

High Throughput Biomass Production. Producing large quantities of high quality and defensibly reproducible cells that have been exposed to specific environmental stressors is critical to high throughput and concomitant analyses using transcriptomics, proteomics, metabolomics, and lipidomics. Culture of *D. vulgaris* is made even more difficult because it is an obligate anaerobe and sulfate reducer. For the past four years, our Genomics:GTL VIMSS project has developed defined media, stock culture handling, scale-up protocols, bioreactors, and cell harvesting protocols to maximize throughput for simultaneous sampling for lipidomics, transcriptomics, proteomics, and metabolomics. All cells for every experiment, for every analysis are within two subcultures of the original ATCC culture of *D. vulgaris*. In the past four years we have produced biomass for 120 (40 in the last year) integrated experiments (oxygen, NaCl, NO₃, NO₂, heat shock, cold shock, pH, Cr, and mutants Fur, Zur, Per, and MP(-)) each with as much as 200 liters of mid-log phase cells (3×10^8 cells/ml). This year new continuous culture extremophile bioreactors were brought online so that six reactors (1-3L) can be operated continuously. This enables U.S. to produce as much as 300L of mid-log phase anaerobe cells in 5 days. In addition, more than 80 adhoc experiments for supportive studies have been done each with 1-6 liters of culture. All cultures, all media components, all protocols, all analyses, all instruments, and all shipping records are completely documented using QA/QC level 1 for every experiment and made available to all investigators on the VIMSS Biofiles database (<http://vimss.lbl.gov/perl/biofiles>). To determine the optimal growth conditions and determine the minimum inhibitory concentration (MIC) of different stressors we adapted plate reader technology using Biolog and Omnilog readers using anaerobic bags and sealed plates. Since each well of the 96-well plate produces an automated growth curve, over more than 200 h, this has enabled U.S. to do more than 10,000 growth curves over the last three years. Since the Omnilog can monitor 50 plates at a time, this allows U.S. to do more than 5,000 growth curves in a year.

Phenotypic Responses. We have generated a large set of phenotypic data that suggest analysis of the strain DePue genome sequence will provide important insights into the acquisition of metal-resistance absent in the closely related strain, *D. vulgaris* Hildenborough. An initial phenotypic characterization of a novel *Desulfovibrio* species isolated from the Hanford demonstration site has been completed and DNA is now being prepared for genome sequencing. We have completed extensive phenotypic comparisons of a large study set of *Desulfovibrio* species (14 different strains), as a prelude to continued comparative studies of fitness and evolution (see other abstract).

Synchrotron FTIR Spectromicroscopy for Real-Time Stress Analysis. This year we further the synchrotron FTIR Spectromicroscopy approach for studying roles of cellular compositions and physiological states during stress and adaptive responses in individual *D. vulgaris* triggered by air-level oxygen. Previously, the FTIR spectroscopy approach has allowed U.S. to detect *in situ* changes in intracellular molecules or molecular structures, and to nondestructively monitor and quantify metabolites produced in response to different stresses. This is because the chemical and structural information of molecules associated with cellular processes inside *D. vulgaris* are contained in each infrared spectrum; thus, one can extract chemical and structural information from each spectrum regarding the physiological conditions of a cell or a group of cells. The improved FTIR spectroscopy approach includes an additional molecular screening procedure, which allows U.S. to rapidly identify individual

D. vulgaris cells that satisfy a targeted chemical composition and physiological state. Such rigidly controlled experimental conditions at chemical and biological levels would improve the reproducibility of experimental results. To date, we have evaluated the new FTIR approach in four different experimental systems using monolayers of wild-type *D. vulgaris* at the early stationary phase. For the first two systems, individual *D. vulgaris* cells of different compositions were maintained anaerobically, which have allowed U.S. to establish baselines for the molecular changes and the timescales associated with cellular processes during anaerobic metabolism. For the remaining two systems, individual *D. vulgaris* cells of different compositions were exposed to air-level oxygen, which have allowed U.S. to establish baselines for the molecular changes and the timescales associated with cellular processes during oxidative stress induced adaptive responses. Many of these results have been confirmed by analysis of microscopy images and biochemical essays. These studies will enable U.S. to do in depth studies of stress mechanisms with the new created mutants from the Functional Genomics Core of the project.

17 ^{GT}

Microarrays + NanoSIMS: Linking Microbial Identity and Function

Jennifer Pett-Ridge^{1*} (pettridge2@llnl.gov), Peter K. Weber,¹ Paul Hoeprich,¹ Philip Banda,¹ **Ian Hutcheon**,¹ Eoin Brodie,² and Gary Andersen²

¹Lawrence Livermore National Laboratory, Livermore, California and ²Lawrence Berkeley National Laboratory, Berkeley, California

Project Goals: We are using a high resolution ion microprobe (Nano Secondary Ion Mass Spectrometer NanoSIMS) to link microbial metabolism to molecular structures and produce a detailed view of how isotopically marked species propagate throughout individual cells. We expose microbes to stable isotope tracers and then map the tracer distribution with the NanoSIMS. Images of cells and microarrays reveal locations of active growth, nutrient fluxes between cells, and functional roles of community members.

In order to predict how microbes may react under given environmental conditions, or be engineered to perform useful functions, it is essential to understand the relationships between their molecular and metabolic profiles. Indeed, our need to understand both the identity and functional capacity of microorganisms is increasing as researchers seek to: a) understand spatial and metabolic relationships within complex microbial communities, b) exploit microbial traits for bioengineered fuel cells and cellulose conversion to biofuels, and c) utilize microbes to remediate contaminated sites.

We are addressing these goals by developing a new methodology, “NanoSIP”, combining the power of re-designed oligonucleotide microarrays with nano-scale secondary ion mass spectrometry (NanoSIMS) analyses in order to link the identity of microbes to their functional roles. Building upon the concept of stable isotope probing (SIP) (Radajewski *et al.* 2000), we are isotopically labeling microbial nucleic acids by growing organisms on ¹³C enriched substrates. When hybridized to a high density oligonucleotide microarray we can use the high spatial resolution and high sensitivity of the NanoSIMS to detect isotopic enrichment in ribosomal RNA fragments identified through fluorescent hybridization to a newly engineered oligonucleotide microarray. This approach will allow U.S. to directly link microbial identity and function.

The NanoSIMS is an imaging secondary ion mass spectrometer with the unprecedented combination of high spatial resolution, high sensitivity and high mass specificity. It has 50 nm lateral resolution and is capable of detecting 1 of every 200 carbon atoms in a sample while excluding isobaric interferences. We have previously used the NanoSIMS to document isotopic and elemental variations in tiny bioparticles such as *Bacillus* spores, bacterial cells and lipid bilayers. Since the spot or feature size on a microarray is typically microns in diameter, and can contain millions of copies of an oligonucleotide probe, the NanoSIMS has the detection capability to resolve array spots labeled with ^{12}C rRNA from those labeled with ^{13}C rRNA.

We are currently in the 'proof-of-concept' phase of method development and are testing the technique using pure cultures of ^{13}C -labelled microbes. Using environmental isolates from a tropical soil, we cultured 2 strains each of fungi, gm (+) bacteria, gm (-) bacteria and actinomycetes with ^{13}C -glucose. Cultures were repeatedly subsampled during exponential phase growth in order to generate a set of samples with a range of isotopic enrichments. We have extracted DNA from these isolates, sequenced the 16S/ITS region and generated 25-mer oligonucleotide probes for each organism. This probe set can be printed onto high density oligonucleotide microarrays using the NimbleGen synthesizer in the LLNL-Livermore Microarray Center (LMAC). The arrays we are using are newly engineered to have a more conductive surface and higher reproducibility relative to traditional glass/silane microarrays. These advances allow U.S. to successfully analyze microarray slides with a nano-secondary ion mass spectrometer (NanoSIMS), generating isotopic and elemental abundance images of the array surface, and indicating which organisms utilized the isotopically labeled substrate. We intend to apply the method to complex microbial communities found in biofilms and soils in the near future.

Reference

1. Radajewski S, Ineson P, Parekh NR & Murrell JC 2000. Stable-isotope probing as a tool in microbial ecology. *Nature* 403: 646-649

18 ^{GTL}

NanoSIMS Analyses of Molybdenum Indicate Nitrogenase and N-Fixation Activity in Diazotrophic Cyanobacteria

Jennifer Pett-Ridge,¹ Juliette Finzi,² **Ian D. Hutcheon**¹ (hutcheon1@llnl.gov), Doug Capone,² and Peter K. Weber^{1*}

¹Chemistry, Materials, and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, California and ²Department of Marine Biology, University of Southern California, Los Angeles, California

Project Goals: We are using a high resolution ion microprobe (Nano Secondary Ion Mass Spectrometer NanoSIMS) to link microbial metabolism to molecular structures and produce a detailed view of how isotopically marked species propagate throughout individual cells. We expose microbes to stable isotope tracers and then map the tracer distribution with the NanoSIMS. Images of cells and microarrays reveal locations of active growth, nutrient fluxes between cells, and functional roles of community members.

Diazotrophic cyanobacteria are capable of both CO_2 and N_2 fixation, yet must separate these two functions because the nitrogenase enzyme critical to N_2 fixation is inhibited by O_2 produced during

photosynthesis. Some lineages, such as *Anabaena oscillarioides*, use specialized cells (heterocysts) to maintain functional segregation. However the mechanism of this segregation is poorly understood in the undifferentiated filamentous *Trichodesmium spp.*, an important component of marine primary production in the tropical and subtropical North Atlantic. While some research on *Trichodesmium IMS101* suggest a temporal segregation of the nitrogen and carbon fixing processes, others indicate nitrogen fixation is spatially isolated in differentiated cells called diazocytes (Fredriksson and Bergman 1997).

In order to isolate the intracellular location of N fixation in both species, we used a combination of TEM, SEM and NanoSIMS analysis to map the distribution of C, N and Mo (a critical nitrogenase co-factor) isotopes in intact cells. NanoSIMS is a powerful *in situ* analysis tool which combines nanometer-scale imaging resolution with the high sensitivity of mass spectrometry. Using cells grown in a $^{13}\text{CO}_2$ and $^{15}\text{N}_2$ enriched atmosphere, our analyses show that heterocysts in *Anabaena* have Mo concentrations four times higher than those of non-N-fixing vegetative cells. Recently fixed N does not accumulate at the site of fixation, but instead is quickly translocated to vegetative cells, presumably to fuel the demands of photosynthesis, storage and cell division.

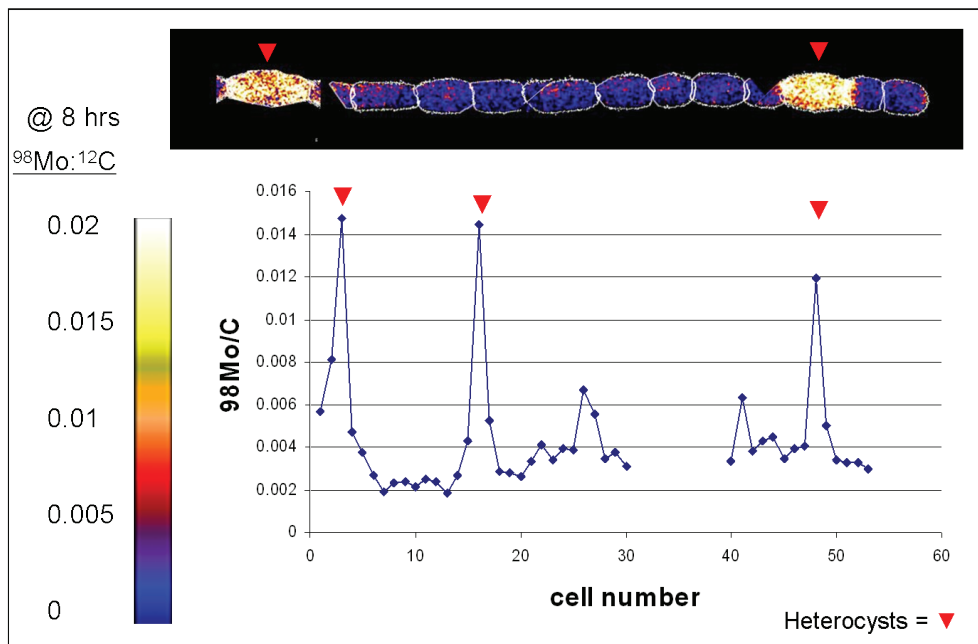


Figure. NanoSIMS image and dataplott of Mo concentrations in a strand of *Anabaena oscillarioides* heterocyst and vegetative cells grown under N-fixation conditions. Brighter colors indicate higher concentrations.

In the non-heterocystous *Trichodesmium IMS101*, Mo is concentrated in sub-regions of individual cells, and is not associated with regions of N storage (cyanophycin granules) which are clearly evident in lateral section TEM images. Average cellular concentrations of Mo increased from $1 (\pm 0.2)$ ppm to $86 (\pm 20)$ ppm during the same early afternoon period when a jump in ^{15}N (and thus N-fixation) was observed. Rare and randomly located cells contained extremely high concentrations of Mo (>2000 ppm).

We suggest that NanoSIMS mapping of metal enzyme co-factors is a powerful method of identifying physiological and morphological characteristics within individual bacterial cells, and could be

used to complement more traditional analyses such as immunogold labeling. Such combinations of NanoSIMS analysis and high resolution microscopy allow isotopic analysis to be linked to morphological features and hold great promise for fine-scale studies of bacteria metabolism.

Reference

1. Fredriksson, C. & Bergman, B. (1997). Ultrastructural characterization of cells specialized for nitrogen fixation in a non-heterocystous cyanobacterium, *Trichodesmium*. *Protoplasma* 197, 76–85

19 ^{GTL}

Application of a Novel Genomics Technology Platform

Mircea Podar,¹ Carl Abulencia,² Don Hutchinson,² Joseph Garcia,² Lauren Hauser,¹ Cheryl Kuske,³ and **Martin Keller**^{1*} (kellerm@ornl.gov)

¹Bioscience Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee; ²Diversa Corporation, San Diego, California; and ³Los Alamos National Laboratory, Los Alamos, New Mexico

Project Goals: The Application of a Novel Genomics Technology Platform combines an isolation method based on fluorescence in situ hybridization (FISH) and cell sorting by flow cytometry, with whole genome amplification (MDA) to obtain a sufficient amount of DNA for sequencing whole genomes of uncultured microorganisms. Soil bacterial representatives of candidate division TM7 were specifically FISH-stained, in suspension, and isolated by flow cytometry. The genomic DNA was subsequently amplified by MDA for the construction of libraries for shotgun sequencing.

Application of cultivation-independent molecular phylogenetic approaches to study microbial communities in the environment led to the discovery of an unexpected genetic diversity and has been followed by an increasing repertoire of environmental genomic tools (expression microarrays, proteomics, and metabolomics). However, the cost and time effort for genomic characterization of most microbial communities through shotgun sequencing is prohibitive due to high microbial diversity and complex distributions of abundance and genome size for the constituent species. To overcome some of these hurdles, we propose a targeted genomic approach. This process combines an isolation method based on fluorescence in situ hybridization (FISH) and cell sorting by flow cytometry, with whole genome amplification (MDA) to obtain a sufficient amount of DNA for sequencing whole genomes of uncultured microorganisms. Soil bacterial representatives of candidate division TM7 were specifically FISH-stained, in suspension, and isolated by flow cytometry (manuscript in preparation). The genomic DNA was subsequently amplified by MDA for the construction of libraries for shotgun sequencing. Based on SSU rRNA sequences, the soil sample studied contained mostly members of the Proteobacteria (35%), Acidobacteria (38%), Gemmatimonadetes (16%) and, at much lower abundance levels (2% or less), representatives of several other phyla. Candidate division TM7 bacteria were among those low-abundance groups, which was appropriate for our goal of targeting a minor constituent of the community for genomic characterization.

We have targeted this approach to the TM7 using specific FISH-staining, in suspension, and isolation of stained bacteria by flow cytometry. A cellular fraction prepared from the soil sample was used for hybridization with a fluorescently labeled oligonucleotide specific for the TM7 phylum (TM7905). Using flow cytometry we detected a small fraction of cells (0.02%) which had a fluorescence level approximately 10 times higher than background based on the unstained control popula-

tion. Fluorescence cells were sorted in pools of various sizes and used for chromosomal amplification. The selected genomic DNA was subsequently amplified by MDA for the creation of libraries for shotgun sequencing of whole genomes. Based on test experiments we determined that five was the fewest number of cells that balanced efficient genomic amplification with low levels of amplification artifacts and chimeric clones.

The MDA-amplified genomic DNA from five sorted cells was used as template for SSU rRNA gene amplification. Among the 69 sequences, 61 (89%) represented a TM7 bacterium. The remaining eight sequences were found to be nearly identical (>99.5%) to SSU ribosomal genes from several environmental *Pseudomonas* isolates including *P. rhodesiae*, an organism isolated from natural mineral waters. These clones may therefore represent an actual *Pseudomonas* cell that was sorted by flow cytometry from the soil sample rather than from contamination of the reagents or instruments.

Sequencing of the amplified DNA has resulted in identification of genes that are from the TM7 genome and will give insights to the functioning of this group. End sequences from 12,000 clones were generated using Sanger-sequencing. After filtering out the low quality and obvious chimeric reads based on Phred/Phrap, approximately 20,000 reads were assembled into contigs using Phrap. Contigs that contained genes with high similarity values to known *Pseudomonas* genes had also elevated GC content (>54%) relative to the bulk of the sequences (<50%) and were filtered out as representing the contaminant. The remaining sequence data, representing ~600kb of, constitutes approximately 15-20% of the TM7 genome, based on statistical distribution of universally present bacterial genes. This genomic data allows for the first time detailed evolutionary analyses of the TM7 phylum as well as insight into the soil TM7 bacterial ecology and metabolism.

Research sponsored by the Genomics:GTL program, Office of Biological and Environmental Research, U.S. Department of Energy Grant No. DE-FG02-04ER63771

20^{GTL}

Genome-Scale Analysis of the Physiological State of *Geobacter* Species During *In Situ* Uranium Bioremediation

Dawn E. Holmes* (dholmes@microbio.umass.edu), Regina A. O'Neil, Milind A. Chavan, Muktak Aklujkar, and **Derek R. Lovley**

Department of Microbiology, University of Massachusetts, Amherst, Massachusetts

Project Goals: 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 goals of Subproject I and II are to determine the genetic potential of the *Geobacteraceae* present in subsurface environments, and to describe genome-wide patterns of gene expression in *Geobacteraceae* species in subsurface environments. 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.

The design of optimal bioremediation strategies for contaminated Department of Energy subsurface sites has been hindered by a lack of information on the *in situ* physiological state of the microorganisms involved in important bioremediation processes and the inability to predict how the microbial

community will respond to different amendments that might be made to promote bioremediation. It is now clear from numerous studies in multiple laboratories at a diversity of sites that stimulation of dissimilatory metal reduction to promote *in situ* bioremediation of uranium and other contaminant metals frequently results in the emergence of *Geobacter* species as the dominant metal-reducing microorganisms. *Geobacter* species are also the predominant organisms involved in the oxidation of various organic contaminants coupled to the reduction of the Fe(III) oxides that are naturally abundant in most subsurface environments.

Further analysis of the *Geobacter* species that predominate in a diversity of subsurface environments revealed a 'subsurface clade' of *Geobacter* species that are consistently the dominant *Geobacter* in a geographic and geochemical diversity of subsurface environments regardless of whether metal or organic contaminants are undergoing bioremediation. This finding greatly simplifies both the identification of gene target sequences for evaluation of the *in situ* metabolic state of *Geobacter* species during groundwater bioremediation and the development of genome-based *in silico* models to predict the metabolic and growth responses of *Geobacter* species under different potential bioremediation strategies.

Three independent and complementary approaches were taken in order to learn more about the genetic potential of the 'subsurface clade' of *Geobacter* species: 1) small insert libraries of genomic DNA extracted directly from the environment were sequenced; 2) genomic DNA was amplified from single cells recovered from the subsurface and sequenced; and 3) strains of *Geobacter* with 16S rRNA gene sequences identical or highly similar to the sequences that predominate during bioremediation were recovered in pure culture, and their genomes were sequenced. Approaches 1 and 2 provide information on the genotypic potential of the *Geobacter* species that predominate during subsurface bioremediation, but are limited in value because: 1) many of the genes identified are of unknown function or if they have an annotated function, their physiological role in *Geobacter* is unclear; 2) it is not possible to predict patterns of gene expression from sequence data alone; and 3) many of the most basic and important phenotypic characteristics, such as optimal conditions for growth, growth yields, etc. cannot yet be predicted from genome sequences alone. Thus, the ability to conduct genome scale analysis of the physiology of environmentally relevant isolates is key to understanding *in situ* physiology and the development of predictive *in silico* models.

Genomes of multiple subsurface *Geobacter* isolates have been sequenced or will be completed shortly. These include: *G. uraniumreducens*, *Geobacter* species M21, and *Geobacter* species M18 recovered from *in situ* uranium bioremediation experiments at the DOE-ERSP field study site in Rifle, CO; strain FRC-32, a *Geobacter* species recovered from an *in situ* uranium bioremediation experiment at the DOE-ERSP Field Research Center at Oak Ridge National Laboratories; *Geobacter* strains Ply1 and Ply4 which were recovered from an acetate-impacted aquifer that serves as an analog for long-term *in situ* uranium bioremediation; and *G. bemiidjensis*, recovered from the Fe(III)-reducing zone of a petroleum-contaminated aquifer. Preliminary results show substantial similarities in the genome sequences of these isolates and the genome sequences obtained from sequencing genomic DNA extracted from the subsurface.

In order to learn more about the physiology of *Geobacter* species growing in subsurface sediments pure cultures of *Geobacter* species were inoculated into sterilized, uranium-contaminated sediments from the ERSP study site in Rifle, CO and the sediments were amended with acetate to simulate conditions during *in situ* uranium bioremediation. Compared to growth on soluble electron acceptors, all three of the species examined, *G. uraniumreducens*, *G. sulfurreducens*, and *G. metallireducens*, had significant increases in transcripts for multiple genes for *c*-type cytochromes, not only during growth in sediment but also in culture medium when Fe(III) or Mn(IV) oxides served as the electron acceptor. There was also increased expression of genes for multicopper oxidase proteins.

The within-strain similarity in gene expression with all three electron acceptors suggests that the mechanisms for electron transfer to subsurface sediment oxides and oxides prepared in the laboratory to simulate sediment oxides are similar. However, between species there were substantial differences in which cytochrome genes were most highly expressed, reflecting the lack of cytochrome gene conservation in *Geobacter* species.

In contrast to the lack of conservation of cytochrome genes, there is high conservation of many other genes across *Geobacter* species and for these genes there were highly similar expression patterns. For example, a number of genes that encode proteins involved in chemotaxis and motility and phosphorus limitation were significantly up-regulated in all of the organisms during growth in sediments or on Fe(III) or Mn(IV) oxides. Furthermore, genes encoding proteins involved in nitrogen fixation, heavy metal stress, and oxidative stress were up-regulated in all three species during growth in sediments, but not when Fe(III) oxide or Mn(IV) oxide were provided as the electron acceptor.

Remarkably, gene expression patterns of pure cultures grown in sediments were highly similar to the *in situ* gene expression of the *Geobacter* species that predominated during *in situ* uranium bioremediation at the Rifle study site. The *Geobacter* species in the groundwater had high transcript levels for genes involved not only in electron transfer to Fe(III) oxides, but also chemotaxis, motility, phosphorus uptake, nitrogen fixation, heavy metal stress, and oxidative stress. These results demonstrate that it is possible to reliably monitor the metabolic state of *Geobacter* species involved in *in situ* uranium bioremediation and suggest that detailed, genome-based physiological studies with pure cultures of environmentally relevant *Geobacter* species can provide insight into the physiology of *Geobacter* species living in subsurface environments. This has important implications for the ability of *in silico* models developed from pure cultures to predict growth and metabolism under different conditions in the subsurface.

Section 3

Protein Production and Characterization

21^{GTL}

High Throughput Selection of Affinity Reagents

Peter Pavlik, Nileena Velappan, Hugh Fisher, Csaba Kiss, Minghua Dai, Emanuele Pesavento, Leslie Chasteen, and **Andrew Bradbury*** (amb@lanl.gov)

Los Alamos National Laboratory, Los Alamos, New Mexico

Project Goals: The goals of this GTL funded project are to implement a high throughput selection and screening system for affinity reagents, with the ability to select against proteins and post-translational modifications. This has required re-engineering of the whole selection and screening process, described in the displayed posters.

Antibodies are the most widely used binding ligands in research. However, they suffer from a number of problems, especially when used in molecular diversity techniques. These include low expression levels, instability and poor cytoplasmic expression, as well the inability to detect binding without the use of secondary reagents. In this project we are creating an integrated in vitro system which will

allow U.S. to select affinity reagents against proteins of interest on a genomic scale. This has involved re-engineering of the whole selection and screening process. Within this context we have developed 1) novel affinity reagents based on fluorescent proteins which resolve many of the problems associated with antibodies (1, 2); 2) new selection methods for such fluorescent affinity reagents (3, 4); 3) high throughput screening systems using flow cytometry (5); 4) eliminated the need for helper phage in phage display selections (6), and 5) shown the application of some of these methods to the selection of antibodies recognizing post-translation modifications independently of sequence context (7)

References

1. Dai, M., Fisher, H.E., Temirov, J., Kiss, C., Phipps, M.E., Pavlik, P., Werner, J.H. and **Bradbury, A.R.M.** (2006) The creation of a novel fluorescent protein by guided consensus engineering, *Prot. Eng. Design Selection* In press
2. Kiss, C., Fisher, H., Pesavento, E., Dai, M., Valero, R., Ovecka, M., Nolan, R., Phipps, L., Velappan, N., Chasteen, L., Martinez, J., Waldo, G.S., Pavlik, P. and **Bradbury, A.R.M.** (2006) Antibody binding loop insertions as diversity elements. *Nuc. Acids Res.*, **34**, e132
3. Dai, M., Pavlik, P. and **Bradbury, A.R.M.** (2007) Using T7 phage display to select GFP based binders, in preparation
4. Velappan, N., Fisher, H., Kiss, C., Chasteen, L., Pavlik, P. and Bradbury, A.R.R. (2007) Optimizing export signals for the phage display of cytoplasmic proteins, in preparation
5. Ayriss, J., Woods, T., **Bradbury, A.R.M.** and Pavlik, P. (2006) High throughput screening of single chain antibodies using multiplexed flow cytometry, *J. Proteomic Res.* In press
6. Chasteen, L., Ayriss, J., Pavlik, P. and **Bradbury, A.R.M.** (2006) Eliminating helper phage from phage display, *Nuc. Acids Res.*, **34**, e145
7. Kehoe, J.W., Velappan, N., Walbolt, M., Rasmussen, J., King, D., Lou, J., Knopp, K., Pavlik, P., Marks, J.D., Bertozzi, C.R., and **Bradbury, A.R.M.** (2006) Using phage display to select antibodies recognizing post-translational modifications independently of sequence context. *Molecular Cellular Proteomics*, in press.

22^{GTL}

Progress on Fluorobodies

Nileena Velappan, Hugh Fisher, Csaba Kiss, Minghua Dai, Emanuele Pesavento, Leslie Chasteen, Peter Pavlik, and **Andrew Bradbury*** (amb@lanl.gov)

Los Alamos National Laboratory, Los Alamos, New Mexico

Project Goals: The goals of this GTL funded project are to implement a high throughput selection and screening system for affinity reagents, with the ability to select against proteins and post-translational modifications. This has required re-engineering of the whole selection and screening process, described in the displayed posters.

Antibodies are the most widely used binding ligands in research. However, they suffer from a number of problems, especially when used in molecular diversity techniques. These include low expression levels, instability and poor cytoplasmic expression, as well the inability to detect binding without the use of secondary reagents. We have developed novel affinity reagents based on fluorescent proteins which resolve many of these problems. However, selection of such affinity reagents remains problematic, because they do not appear to be very well displayed on phage. In addition to using a cytoplasmic phage, T7 (1), we have undertaken an examination of the ability of three different translocation

pathways (Sec, SRP, TAT) used by *E. coli* to transfer proteins into the periplasm, to incorporate GFP and modified GFP into phage particles, by placing different leaders upstream of GFP (2). We find that while superfolder GFP is efficiently translocated and incorporated into phage in a functional manner with all three leaders, GFP which has been modified, by the insertion of a binding loop, for example, can only be effectively incorporated into phage using TAT based leaders. This provides an effective phage display platform with which to select fluorescent protein based affinity reagents.

Additional data on the success in selecting affinity reagents with intrinsic fluorescence will be presented.

References

1. Dai, M., Pavlik, P. and **Bradbury, A.R.M.** (2007) Using T7 phage display to select GFP based binders, in preparation
2. Velappan, N., Fisher, H., Kiss, C., Chasteen, L., Pavlik, P. and Bradbury, A.R.R. (2007) Optimizing export signals for the phage display of cytoplasmic proteins, in preparation

23 ^{GTL}

High Throughput Screening of Affinity Reagents: Eliminating Helper Phage from Phage Display by the Use of Helper Plasmids

Leslie Chasteen, Joanne Ayriss, Nileena Velappan, Peter Pavlik, and **Andrew Bradbury*** (amb@lanl.gov)

Los Alamos National Laboratory, Los Alamos, New Mexico

Project Goals: The goal of this GTL funded project is the implementation of a high throughput affinity reagent selection and screening program against proteins and post-translational modifications. This has required re-engineering many aspects of the selection and screening process described in the displayed posters.

Phage display is a relatively straightforward technology used to generate binding ligands against a vast number of different targets, involvings the display of proteins or peptides, as coat protein fusions, on the surface of a phage or phagemid particles. However, the need to use helper phage for the replication and assembly of phagemid particles, during library production and biopanning, has prevented full automation of the selection process. Helper phage are added at precise periods of bacterial growth, and it is impossible to avoid contamination of the phage output with helper phage. We have eliminated the need to add helper phage by using “bacterial packaging cell lines” that provide the same functions. These cell lines contain M13 based helper plasmids that express phage packaging proteins which assemble phagemid particles as efficiently as helper phage, but without helper phage contamination; resulting in genetically pure phagemid particle preparations. Furthermore, by using constructs differing in the form of gene 3 that they contain, we have shown that the display, from a single library, can be modulated between monovalent (phagemid-like) to multivalent display (phage-like) without any further engineering. These packaging cells eliminate the use of helper phage from phagemid based selection protocols; reducing the amount of technical preparation, facilitating automation, optimizing selections by matching display levels to diversity, and effectively using the packaged phagemid particles as means to transfer genetic information at an efficiency approaching 100%.

By eliminating the need to add helper phage at precise stages of bacterial growth, and avoiding contamination of the output phagemid particles with helper phage, the use of these cells rather than helper phage will considerably facilitate automation of phage display selection.

Reference

1. Chasteen, L., Ayriss, J., Pavlik, P. and **Bradbury, A.R.M.** (2006) Eliminating helper phage from phage display, *Nuc. Acids Res.*, **34**, e145

24 [—]_{GTL}

Selecting Affinity Reagents which Recognize Specific Post-Translational Modifications Independently of Sequence Context: The Sulfotyrosine Example

John Kehoe,³ Jytte Rasmussen,² Monica Walbolt,² Jianlong Lou,⁴ James D. Marks,⁴ Peter Pavlik,¹ Carolyn Bertozzi,² and **Andrew Bradbury**^{1*} (amb@lanl.gov)

¹Los Alamos National Laboratory, Los Alamos, New Mexico; ²University of California, Berkeley, California; ³Centocor, Horsham, Pennsylvania; and ⁴University of California, San Francisco, California

Project Goals: The goal of this GTL funded project is the implementation of a high throughput affinity reagent selection and screening program against proteins and post-translational modifications. This has required re-engineering many aspects of the selection and screening process described in the displayed posters.

Many cellular activities are controlled by post-translational modifications (PTMs), the study of which is hampered by the lack of specific reagents. The small size and ubiquity of such modifications makes the use of immunization to derive global antibodies able to recognize them independently of context extremely difficult. Here we demonstrate how phage display can be used to generate such specific reagents, using sulfotyrosine as an example. This modification is important in many extracellular protein-protein interaction, including the interaction of some chemokines with their receptors, and HIV infection.

We designed a number of different selection strategies, using peptides containing the sulfotyrosine modification as positive selectors in the presence of an excess of the non-modified peptide as blocking agent. We screened almost eight thousand clones after two or three rounds of selection and identified a single scFv able to recognize tyrosine sulfate in multiple sequence contexts. Further analysis shows that this scFv is also able to recognize naturally sulfated proteins in a sulfation dependent fashion, and its binding could be inhibited by soluble tyrosine sulfate, but not tyrosine or tyrosine phosphate, providing an excellent way to control for the specificity of binding. This scFv was converted into a full length IgG and into an scFv-AP fusion, both of which increased the stability. This antibody has been distributed to a number of different groups which have used it successfully, some results of which will be presented.

It has proved to be extremely difficult to generate antibodies able to recognize post-translational modifications independently of sequence context by immunization, with antibodies against phosphotyrosine being the only well documented example. The use of phage display, as described here,

provides proof of principle for the use of this technology to develop similar reagents against other post-translational modifications.

Reference

1. Kehoe, J.W., Velappan, N., Walbolt, M., Rasmussen, J., King, D., Lou, J., Knopp, K., Pavlik, P., Marks, J.D., Bertozzi, C.R., and **Bradbury, A.R.M.** (2006) Using phage display to select antibodies recognizing post-translational modifications independently of sequence context. *Molecular Cellular Proteomics*, in press.

25^{GTL}

A Total Chemical Synthesis Approach to Protein Structure and Function

Stephen Kent* (skent@uchicago.edu), Duhee Bang, Thomas Durek, Zachary Gates, Erik Johnson, Brad Pentelute, and Vladimir Torbeev

Institute for Biophysical Dynamics, University of Chicago, Chicago, Illinois

Project Goals: Our goal is to address the known limitations of chemical protein synthesis, based on our intimate understanding of the current state of the art. Emphasis will be on the development of simple methods using low cost hardware wherever possible. In this way, we will develop a practical chemical protein synthesis technology applicable to the rapid preparation of milligram amounts of small and integral membrane protein targets based on predicted gene sequence data. We will prototype the application of these methods to selected proteins of the model organism *Shewanella oneidensis* and proteins from a range of other sources, to illustrate potential application of chemical protein synthesis to validating the annotation of microbial genomes. The resulting knowledge will form the basis for future high throughput, parallel chemical synthesis of protein molecules that are difficult to prepare by recombinant DNA expression methods.

Microbial ‘proteins’ are being discovered at an accelerating pace, thanks to the successes of genome sequencing. Using advanced bioinformatics, in the past ten years many tens-of-thousands of predicted proteins have been added to the databases. Our next challenge is to validate the annotation of microbial genomes in terms of the mature protein translation products and their putative functions. One powerful, if underappreciated, way of doing this is *total chemical protein synthesis* – the use of organic chemistry to construct the predicted polypeptide chain, followed by folding of the synthetic polypeptide to give the unique, defined tertiary structure of the protein molecule. The synthetic product is then used to confirm the predicted biochemical function. Proteins prepared by total chemical synthesis have proved to be especially useful for determining the three-dimensional structure of the protein molecule by high resolution X-ray crystallography. Subsequent to these baseline observations, total chemical synthesis provides an efficient and versatile tool for elucidating in unique ways the molecular basis of protein function. For example, variant synthetic proteins corresponding to predicted post-translational modifications (e.g. phosphorylation; lipidation) can be readily prepared in defined, metabolically stable forms and then used to explore the effects on biochemical function (Ref SEP). Chemical protein synthesis is uniquely enabling for the application to proteins of advanced biophysical methods: e.g. selective labeling with nmr probe nuclei; single molecule fluorescence studies.

Over the past ten years, many hundreds of protein molecules have been successfully prepared by total chemical synthesis, typically in multiple tens-of-milligram amounts of high purity, correctly folded product. Total synthesis is particularly suited to the efficient preparation of small proteins (less than ~100 residues), Cys-rich proteins, and integral membrane (IM) proteins. Modern total protein synthesis has evolved from the 'chemical ligation' methods introduced in the mid-1990s (Refs.). Unprotected synthetic peptide segments, spanning the amino acid sequence of the target polypeptide chain, are covalently joined to one another in quantitative yield, without enzymes, by chemoselective reaction of unique, mutually reactive functional groups on each segment. Native chemical ligation ('NCL'), thioester-mediated chemoselective reaction at Cys residues, is the most robust and useful of ligation chemistry developed to date. Chemical protein synthesis is straightforward and the outcome quite predictable; the challenge for most laboratories is making the peptide-thioester building blocks.

We will present a series of case studies from our ongoing work, to illustrate the current capabilities of chemical protein synthesis and some of its applications. These case studies include:

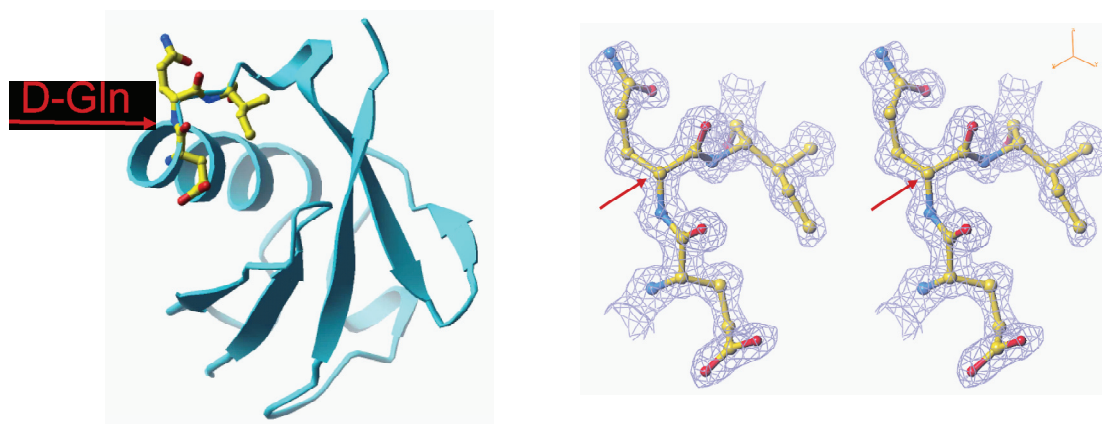


Figure 1. Synthesis and X-ray structures of ubiquitin and D-amino acid ubiquitin analogues.

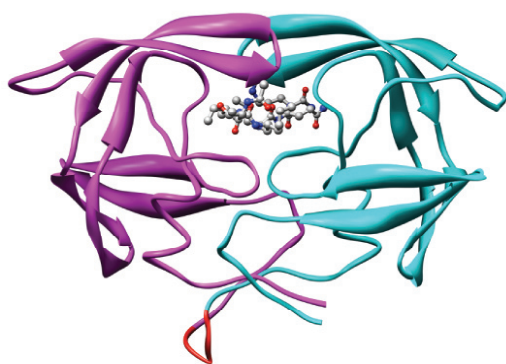


Figure 2. Convergent synthesis and X-ray structure of a 203 amino acid residue 'covalent dimer' form of the HIV-1 protease

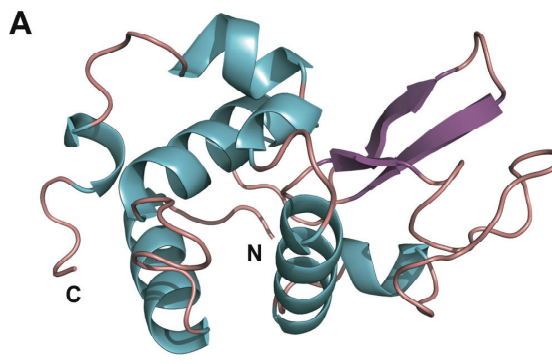


Figure 3. Convergent synthesis and 1.04Å X-ray structure (above) of human lysozyme.

Other topics that will be discussed include: high efficiency synthesis of a series of transmembrane peptide-thioesters spanning the sequence of the protein diacylglycerol kinase, an integral membrane enzyme; 'kinetically controlled ligation' for the fully convergent synthesis of protein molecules; and, convergent synthesis of the plant protein crambin.

We will describe recent innovations that extend the range of targets to which chemical protein synthesis can be applied. Future developments will include the high volume production of peptide-thioesters (modified chemistry, automated/parallel synthesis & purification), and high throughput micro-scale chemical protein synthesis using laboratory automation and microfluidics. Such improvements will enable the broad application of total chemical protein synthesis to the annotation of microbial genomes.

References

1. Constructing proteins by dovetailing unprotected synthetic peptides: backbone engineered HIV protease. M. Schnölzer, S. Kent *Science*, **256**, 221-225 (1992)
2. Synthesis of proteins by native chemical ligation. Philip E. Dawson, Tom W. Muir, Ian Clark-Lewis, Stephen B.H. Kent, *Science*, **266**, 776-779 (1994)
3. Synthesis of native proteins by chemical ligation. Dawson, P.E., Kent S.B.H. *Ann. Rev. Biochem.* **69**, 925-962 (2000)
4. Design and chemical synthesis of a homogeneous polymer-modified erythropoiesis protein. Gerd G. Kochendoerfer, et al., *Science*, **299**, 884-887 (2003)
5. Dissecting the energetics of protein α -helix C-cap termination through chemical protein synthesis. Duhee Bang, Alexey V. Gribenko, Valentina Tereshko, Anthony A. Kossiakoff, Stephen B. Kent*, George I. Makhatadze*, *Nature Chemical Biology*, **2**, 139-43 (2006)
6. Towards the total chemical synthesis of integral membrane proteins: a general method for the synthesis of hydrophobic peptide- α thioester building blocks. E.C.B. Johnson, S.B.H. Kent, *Tetrahedron Letters*, submitted (2006)
7. Kinetically-controlled ligation for the convergent chemical synthesis of proteins. Duhee Bang, Brad Pentelute, Stephen B.H. Kent, *Angew Chem Int Ed Engl.*, **45**, 3985-3988 (2006)
8. Total synthesis of proteins by convergent chemical ligation of unprotected peptides. T. Durek, D. J. Boerema, Z. P. Gates, S. Liu, B. L. Pentelute, V.Yu. Torbeev, Stephen B. H. Kent, submitted (2006)
9. Convergent chemical synthesis and high resolution X-ray structure of human lysozyme. Thomas Durek, Vladimir Yu. Torbeev, Stephen B. H. Kent, submitted (2006)
10. Convergent chemical synthesis and crystal structure of a 203 amino acid 'covalent dimer' HIV-1 protease enzyme molecule. Vladimir Yu. Torbeev, Stephen B. H. Kent, *Angew Chem Int Ed Engl*, accepted for publication (2006)

26 ^{GTL}

A Combined Informatics and Experimental Strategy for Improving Protein Expression

Osnat Herzberg, **John Moulton*** (moulton@umbi.umd.edu), Fred Schwarz, and Harold Smith

Center for Advanced Research in Biotechnology, Rockville, Maryland

Project Goals: Improved success rates for recombinant protein expression are critical to many aspects of the Genomics:GTL program. This project is focused on determining which factors determine whether or not soluble protein is produced in *E. coli*, and using the results to develop a set informatics and experimental strategies for improving expression results. A three pronged strategy is used: experimental determination of the stability and folding properties of insoluble versus soluble expressers, examination of the cellular response to soluble and insoluble expressers, and informatics and computer modeling.

Improved success rates for recombinant protein expression are critical to many aspects of the Genomics:GTL program. This project is focused on determining which factors determine whether or not soluble protein is produced in *E. coli*, and using the results to develop a set informatics and experimental strategies for improving expression results. A three pronged strategy is used: experimental determination of the stability and folding properties of insoluble versus soluble expressers, examination of the cellular response to soluble and insoluble expressers, and informatics and computer modeling.

Informatics methods have been used to examine a wide range of factors potentially affecting soluble expression, including protein family size, native expression level, low complexity sequence, open reading frame validity, amyloid propensity and inherent disorder. Of these, the most significant ones affecting expression outcome are native expression level, family size, and inherent disorder. Surprisingly, a relatively high fraction of disorder is also found to be a characteristic of 'singletons'. We are currently experimenting with machine learning methods, incorporating all of the above factors, as a means of predicting soluble expression.

Transcriptional profiling has revealed a reproducible pattern of gene expression in response to the accumulation of insoluble recombinant protein. The transcriptome partially overlaps those observed during heat shock induction or culture saturation, indicative of regulation, in part, by sigma factors 32 and 38 (encoded by *rpoH* and *rpoS*, respectively). We have used this information to develop a GFP reporter plasmid for insoluble protein accumulation, and identified sigma38 as a key regulator of its expression. Currently, efforts are underway to engineer the promoter of the reporter plasmid to decrease background GFP expression while retaining the ability to discriminate between soluble vs. insoluble protein accumulation.

Protein stability measurements on a set of 12 bacterial proteins have been performed using differential scanning calorimetry and chemical denaturation with guanidine hydrochloride. The data from the two methods are in good agreement, and confirm the earlier finding of that stability is not major factor in determining soluble expression. Work is now underway to investigate the folding properties of these proteins.

This project is supported by Genomics:GTL award DE-FG02-04ER63787.

27 [—]_{GTL}

Structural and Functional Characterization of a Periplasmic Sensor Domain from *Geobacter sulfurreducens* Chemotaxis Protein: A Novel Structure from a Family of Sensors in *Geobacteraceae*

P. Raj Pokkuluri,¹ Yuri Y. Londer,¹ Norma Duke,¹ Stephan Wood,¹ Miguel Pessanha,² Teresa Catarino,³ Carlos A. Salgueiro,² and **Marianne Schiffer**^{1*} (mschiffer@anl.gov)

¹Biosciences Division, Argonne National Laboratory, Argonne, Illinois; ²Requimte, CQFB, Dep. Quimica, FCT-UNL, Caparica, Portugal; and ³Instituto de Tecnologia Quimica e Biologica, UNL, Oeiras, Portugal.

Project Goals: As sub-project of GTL grant "Genome-based models to optimize in situ bioremediation of uranium and harvesting electrical energy from waste organic matter, Derek Lovley (PI)" our goals are to analyze selected proteins to understand their function in the cell. This

includes modeling of structures based on their amino acid sequences, determination of their structures, and the functional interpretation of the structures, such as active sites and surface properties.

Geobacter sulfurreducens encodes over 100 cytochromes containing *c*-type hemes. *G. sulfurreducens* also has one of the largest numbers of proteins annotated as parts of the two-component signal transduction and/or chemotaxis pathways. Ten of the signal transducers have a periplasmic sensor domain which are homologous to each other, and contain sequence signature for *c*-type hemes (1). Two of these domains from methyl-accepting chemotaxis proteins encoded by genes GSU0582 and GSU0935, were expressed in *E. coli* co-transformed with the plasmid bearing cytochrome *c* maturation genes. The domains have about 135 residues, 40% of which are identical.

The heme groups in both proteins are five coordinated in their oxidized state and six-coordinated in their reduced state. The binding patterns for NO, and CO were determined by UV-Vis and NMR spectroscopies. Both proteins bind NO in their oxidized and reduced forms. CO only binds in the reduced state, replacing the endogenous sixth axial ligand of the heme. UV-Vis spectroscopy showed that imidazole is bound only in the oxidized state and it forms the sixth ligand to the heme. The ligand switch upon binding CO suggests a conformational change in the protein which could be a mechanism for signal transduction by these molecules. Both domains have a negative reduction potential: -169mV and -264mV for GSU0582 and GSU0935, respectively. The 95mV difference between their redox potentials suggests different biological functions for these domains.

Remarkably, although the UV-Vis spectra indicate that the heme of these domains is similar to that of cytochrome *c*', their structure is predicted by the program 3D-PSSM (2) to be homologous to CitAP, the periplasmic citrate-binding PAS domain of sensor kinase that does not contain heme (3). We now crystallized the sensor domain from GSU0935 and determined its structure *de novo* using the anomalous dispersion of the iron atom of the heme at the Structural Biology Center beam line of the APS. As predicted by the program 3D-PSSM, the structure is indeed homologous to CitAP. Interestingly, only 13% of the residues is identical between CitAP and sensor domain of GSU0935; the heme binding site is found to be located in an inserted segment as predicted (1). The crystallographic refinement is in progress; details of the structure will be discussed.

The structure of sensor domain of GSU0935 is the first structure of a PAS domain that contains a covalently bound heme. This sensor domain from chemotaxis protein GSU0935 represents a previously unreported family of PAS-type periplasmic sensor domains; these domains could be part of an important mechanism for sensing redox potential or small ligands in the periplasm. Homologs to the sensor domains we identified in *G. sulfurreducens* are observed in various bacteria although they occur in larger numbers in the *Geobacteraceae*.

References

1. Londer YY, Dementieva IS, D'Ausilio CA, Pokkuluri PR & Schiffer M (2006) Characterization of a *c*-type heme containing PAS sensor domain from *Geobacter sulfurreducens* representing a novel family of periplasmic sensors in *Geobacteraceae* and other bacteria. *FEMS Microbiol Lett* **258**: 173-181.
2. Kelley LA, MacCallum RM & Sternberg MJE (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J Mol Biol* **299**: 499-520.
3. Reinelt S, Hofmann E, Gerharz T, Bott M & Madden DR (2003) The structure of the periplasmic ligand-binding domain of the sensor kinase CitA reveals the first extracellular PAS domain. *J Biol Chem* **278**: 39189-39196.

28 [—]_{GTL}**High-Throughput Production and Analyses of Purified Proteins**

F. William Studier^{1*} (studier@bnl.gov), John C. Sutherland,^{1,2} Lisa M. Miller,³ Hui Zhong,³ and Lin Yang³

¹Biology Department, Brookhaven National Laboratory, Upton, New York; ²East Carolina University, Greenville, North Carolina; and ³National Synchrotron Light Source, Brookhaven National Laboratory, Upton, New York

Project Goals: The work is aimed at improving the efficiency of high-throughput protein production from cloned coding sequences and developing a capacity for high-throughput biophysical characterization of the proteins obtained.

This work is aimed at improving the efficiency of high-throughput protein production from cloned coding sequences and developing a capacity for high-throughput biophysical characterization of the proteins obtained. Proteins are produced in the T7 expression system in *Escherichia coli*, which is capable of expressing a wide range of proteins. New vector/host combinations, combined with non-inducing and auto-inducing growth media, provide stable, reliable and convenient expression, even for proteins that are highly toxic to the host and cannot be maintained in the usual pET vectors.

Proteins produced from clones are often improperly folded or insoluble. Many such proteins can be solubilized and properly folded, whereas others appear soluble but remain aggregated or improperly folded. As high-throughput production of purified proteins becomes implemented in GTL projects and facilities, reliable analyses of the state of purified proteins will become increasingly important for quality assurance and to contribute functional information. Beam lines at the National Synchrotron Light Source analyze proteins by small-angle X-ray scattering (SAXS) to determine size and shape, X-ray fluorescence microprobe to identify bound metals, and Fourier transform infrared (FTIR) and UV circular dichroism (CD) spectroscopy to assess secondary structure and possible intermolecular orientation. A liquid-handling robot for automated loading of samples from 96-well plates for analysis at each of these stations has been built and implemented with purified proteins. These data are being used as a training set for neural network analysis of new proteins, to determine whether they are folded properly, obtain information on dynamics and stability, and provide an approximate structure classification.

This project is supported by the Office of Biological and Environmental Research of the Department of Energy. Work on auto-induction and vector development also receives support from the Protein Structure Initiative of the National Institute of General Medical Sciences of NIH, as part of the New York Structural Genomics Research Consortium.

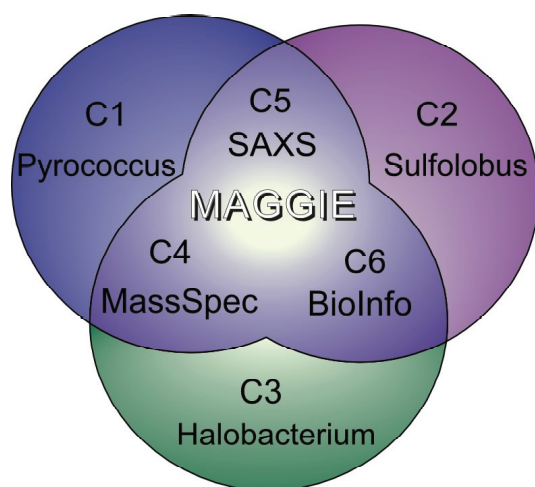
Section 4

Molecular Interactions

29 ^{GTL}**Molecular Assemblies, Genes, and Genomics Integrated Efficiently: MAGGIE****John A. Tainer*** (jat@scripps.edu)

Life Science Division, Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California

Project Goals: MAGGIE integrates an interdisciplinary team at Lawrence Berkeley National Lab with researchers at The Scripps Research Institute, the University of Georgia, the University of California Berkeley, and the Institute for Systems Biology into a unified Genomics GTL program. Major overall goals are 1) to facilitate instrument and technology development and optimizations through cross-disciplinary collaborations, 2) to comprehensively characterize complex molecular machines including protein complexes (PCs) and modified proteins (MPs) and 3) to provide critical enabling technologies and a prototypical map of PCs and MPs for the GTL Program.



MAGGIE integrates an interdisciplinary team at Lawrence Berkeley National Lab with researchers at The Scripps Research Institute, the University of Georgia, the University of California Berkeley, and the Institute for Systems Biology into a unified Genomics GTL program. Major overall goals are 1) to facilitate instrument and technology development and optimizations through cross-disciplinary collaborations, 2) to comprehensively characterize complex molecular machines including protein complexes (PCs) and modified proteins (MPs) and 3) to provide critical enabling technologies and a prototypical map of PCs and MPs for the GTL Program.

MAGGIE focuses on providing an integrated, multi-disciplinary program and synchrotron facilities at the Advanced Light Source (ALS) to achieve efficient key technologies and databases for the molecular-level understanding of the dynamic macromolecular machines that underlie all of microbial cell biology. Together the six MAGGIE Component Subprojects have complementary and synergistic capabilities that unite and leverage the biophysical strengths at LBNL and the ALS with those of top university and research institutes. The Program management and data sharing is promoting synergistic investigator interactions to provide interdisciplinary expertise and scientific critical mass to meet the emerging experimental challenges. Although a new program, we have already had substantial progress as shown on our website: <http://masspec.scripps.edu/MAGGIE/index.php> and in our publications (see below).

MAGGIE is moving to meet the challenges posed by comprehensive characterizations of molecular machines by combining the advantages of specific microbial systems with those of advanced technologies. We highlight 7 initial accomplishments for the overall program: 1) the *Pyrococcus* system is providing PCs and MPs from native biomass, 2) the *Sulfolobus* system is providing genetics for tagged complexes, 3) the *Halobacterium* system is providing extensive system biology results and capabilities, 4) novel developments in high throughput mass spectrometry promise to make large impacts on the research community, 5) the SIBLYS beamline and SAXS facilities are now working as unique and productive world class facilities to visualize PCs and MPs in solution, 6) graph theory is providing characterizations of protein module interactions using cliques, and 7) GAGGLE software is providing a superb technology for communications across multiple databases.

Publications from MAGGIE funding

Facciotti M.T., Pan M., Kaur A., Vuthoori M., Reiss D.J., Bonneau R., Shannon P., Srivastava A., Donahoe S.M., Hood L., Baliga N.S. "Structure of a general transcription factor specified global gene regulatory network," submitted, 2006

Schmid A.K., Reiss D.J., Kaur A., Pan M., King N., Hohmann L., Baliga N.S. "Tracking transcriptome and proteome dynamics during oxic/anoxic transitions in cellular physiology," submitted, 2006

Whitehead K., Kish A., Pan M., Kaur A., Reiss D.J., King N., Hohmann L., DiRuggiero J., Baliga N.S. "An integrated systems approach for understanding cellular responses to gamma radiation," *Mol Syst Biol*, 2: 47, 2006.

Schmid A., Baliga N. "Prokaryotic Systems Biology," *In Cell Engineering*, El-Rubeai, M. (ed): Springer, 5, 2006.

Bonneau R., Reiss D.J., Shannon P., Facciotti M., Hood L., Baliga N.S., Thorsson V. "The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo," *Genome Biol*, 7: R36, 2006.

Reiss D.J., Baliga N.S., Bonneau R. "Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks," *BMC Bioinformatics*, 2006, 7: 280.

Shannon P., Reiss D.J., Bonneau R., Baliga N.S. "Gaggle: An open-source software system for integrating bioinformatics software and data sources," *BMC Bioinformatics*, 7: 176, 2006.

Kaur A., Pan M., Meislin M., Facciotti M.T., El-Geweley R., Baliga N.S. "A systems view of haloarchaeal strategies to withstand stress from transition metals," *Genome Res*, 16: 841-854, 2006.

Want E.J., Nordstrom A., Morita H., Siuzdak G. "From Exogenous to Endogenous: The Inevitable Imprint of Mass Spectrometry in Metabolomics," submitted, 2006.

Go E.P., Wikoff W., Shen Z., O'Maille G., Morita H., Conrads T.P., Nordstrom A., Trauger S.A., Uritboonthai W., Lucas D., Chan K.C., Veenstra T.D., Lewicki H., Oldstone M.B., Schneemann A., Siuzdak G. "Mass Spectrometry Reveals Specific and Global Molecular Transformations during Viral Infection," *Journal of Proteome Research*, in press, 2006.

Shen Z., Want E.J., Chen W., Keating W., Nussbaumer W., Moore R., Gentle T.M., Siuzdak G. "Sepsis Plasma Protein Profiling with Immunodepletion, Three-Dimensional Liquid Chromatography Tandem Mass Spectrometry and Spectrum Counting," *Journal of Proteome Research*, in press, 2006.

- Northern T.R., Northern M.T., Nordstrom A., Uritboonthai W., Turner K., Siuzdak G. "A Surface Rearrangement Mechanism for Desorption/Ionization on Porous Silicon," submitted, 2006.
- O'Maille G., Hoang L., Nordstrom A., Go E.P., Qin C., Siuzdak G. "Enhanced Metabolite Profiling via Chemical Derivatization and Isotope Labeling," submitted, 2006.
- Nordstrom A., O'Maille G., Qin C., Siuzdak G. "Non-linear Data Alignment for UPLC-MS and HPLC-MS based Metabolomics: Quantitative Analysis of Endogenous and Exogenous Metabolites in Human Serum," *Analytical Chemistry*, 78, 7289-3295, 2006.
- Go E.P., Uritboonthai W., Apon J.A., Trauger S.A., Nordstrom A., O'Maille G., Brittain S., Peters E.C., Siuzdak G. "Fluorous Affinity Tags for Selective Metabolite and Peptide Capture and Mass Detection," submitted, 2006.
- Want E.J., Smith C., Siuzdak G. "Phospholipid Capture Combined with Non-Linear Chromatographic Correction for Improved Metabolite Profiling," *Metabolomics*, in press, 2006.
- Fan L., Arvai A., , Cooper P.K., Iwai S., Hanaoka F., Tainer J.A. "Conserved XPB Core Structure and Motifs for DNA Unwinding: Implications for Pathway Selection of Transcription or Excision Repair," *Molecular Cell*, 22: 27-37, 2006.
- Pascal J. M., Tsodikov O.V., Hura G.L., Song W., Cotner E.A., Classen S., Tomkinson A.E., Tainer J.A., Ellenberger T. "A flexible interface between DNA ligase and a heterotrimeric sliding clamp supports conformational switching and efficient ligation of DNA," *Molecular Cell*, 24:279-91, 2006.
- Tsutakawa S.E., Hura G.L., Frankel K.A., Cooper P.K., Tainer J.A. "Structural analysis of flexible proteins in solution by small angle X-ray scattering combined with crystallography," *J. Structural Biology*, in press, 2006.
- Chris H.Q. Ding, Xiaofeng He, and Stephen R. Holbrook, "Transitive closure and metric inequality of weighted graphs – detecting protein interaction modules using cliques," *Int. J. Data Mining and Bioinformatics Vol.1, No.2*. 2006.
- C. Wang, C. Ding, R.F. Meraz, and S.R. Holbrook, "PSoL: a positive sample only learning algorithm for finding non-coding RNA genes," *Bioinformatics*, 22:2590-2596, 2006.
- Chris Ding, Ya Zhang, and Stephen Holbrook, "Biclustering Protein Complex Interactions with a Biclique Finding Algorithm," 2006 IEEE International Conference on Data Mining, IEEE Computer Society Press (in press), 2006.
- Chunlin Wang, Chris H.Q. Ding & Stephen R. Holbrook, "Anatomy of the Yeast Protein Interaction Network by Hierarchical Decomposition," (Submitted to *Nature Biotechnology*), 2006.
- Ya Zhang, Chris Ding and Stephen Holbrook, "Simultaneously Relating Domains and Protein complexes through Biclique Discovery," (Submitted to *Bioinformatics*), 2006.

30 ^{GTL}

The MAGGIE Project: Identification and Characterization of Native Protein Complexes and Modified Proteins from *Pyrococcus furiosus*

Angeli Lal Menon^{1*} (almenon@uga.edu), Farris L. Poole II,¹ Aleksandar Cvetkovic,¹ Saratchandra Shanmukh,¹ Joseph Scott,¹ Francis E. Jenney Jr.,¹ Sunia Trauger,^{2,3} Ewa Kalisiak,^{2,3} Gary Siuzdak,^{2,3} Greg Hura,³ John A. Tainer,³ and **Michael W. W. Adams**¹

¹Department of Biochemistry and Molecular Biology, University of Georgia, Athens, Georgia; ²Center for Mass Spectrometry, The Scripps Research Institute, La Jolla, California; and ³Department of Biochemistry and Molecular Biology, The Scripps Research Institute, La Jolla, California

Project Goals: Our goals are to (a) identify native multiprotein complexes (PCs) and modified proteins (MPs), such as those containing organic and/or inorganic cofactors, using native bio-mass of a model hyperthermophilic organism, *Pyrococcus furiosus*, by mass spectrometry in collab-

oration with Gary Suizdak, (b) to provide native samples of the more abundant PCs and MPs for characterization by small angle X-ray scattering (SAXS) in collaboration with John Tainer, (c) to use bioinformatic approaches to validate and define PCs and MPs for multiple ORF expression in collaboration with Steve Yannone, Nitin Baliga and Steve Holbrook, (d) to produce recombinant PCs on analytical and preparative scales for structural characterization in collaboration with John Tainer, (e) to design homologous and heterologous genetic approaches for the production and characterization of PCs in collaboration with Steve Yannone, Nitin Baliga (f) to design and evaluate generic protocols for PC and MP protein production in other prokaryotic systems of DOE interest, with Steve Yannone, Nitin Baliga and Steve Holbrook.

Most cellular processes are carried out by dynamic molecular machines or large protein complexes (PCs), and many of which include post-translationally-modified proteins (MPs), such as those containing organic and/or inorganic cofactors. Despite the fact that most cellular proteins exist in the form of stable or transient PCs, their composition and the ORFs that encode the components of these complexes are largely unknown. They cannot be predicted from bioinformatics analyses. In addition, no well defined techniques are currently available to unequivocally identify PCs or MPs and their individual components. Some of these issues can be resolved by determining the identity of PCs and MPs found in native proteomes. We are using the archaeon, *Pyrococcus furiosus*, a hyperthermophile that grows optimally near 100°C, as the model organism. By analyzing the native proteome at ambient temperatures, close to 80°C below the optimal physiological and growth temperature, the goal is to capture both stable and dynamic/transient protein complexes for identification, purification, and molecular and functional characterization.

Large scale fractionation of native *P. furiosus* biomass is being accomplished using non-denaturing, column chromatography techniques. Samples from the column fractions are being analyzed by native and denaturing PAGE, mass spectrometry (nano LC-ESI-MS/MS and MALDI-MS) and metal analyses (colorimetric and ICP-MS) to identify PCs and MPs and to determine their relative abundance in the native biomass. The more abundant PCs and MPs obtained from native biomass fractionation are being directly analyzed by Multiple Angle Light Scattering (MALS), Dynamic Light Scattering (DLS) and Small Angle X-ray Scattering (SAXS) to provide information on purity, native complex mass and subunit stoichiometry. Purified, abundant native complexes are also being used for structural characterization. The less abundant PCs and MPs and their individual components are being produced using recombinant gene expression and purification based on bioinformatic predictions and data from the native biomass analyses. The recombinant portion of the project takes advantage of the pre-existing infrastructure developed for a previous structural genomics effort with *P. furiosus*. In a preliminary pilot study almost 600 proteins were identified in fractions eluted from the first chromatographic separation of the cytoplasmic fraction from native *P. furiosus* biomass. Of these, 108 were proposed to be part of 45 potential heteromeric complexes in high abundance according to their elution behavior. A total of 29 of the 45 were previously uncharacterized, consisting of predominantly conserved hypothetical proteins, and not predicted to encode PCs. Approximately half of the fractions from the first chromatography step were subsequently fractionated by a total of 16 additional chromatography steps yielding almost 1000 distinct fractions. The nature of the PCs and MPs (particularly metal-containing proteins) that were identified and purified in this pilot study of native biomass will be described.

31 ^{GTL}**The MAGGIE Project: Production and Isolation of Tagged Native/ Recombinant Multiprotein Complexes and Modified Proteins from Hyperthermophilic *Sulfolobus solfataricus***

Denise Munoz,¹ Jill Fuss,¹ Kenneth Stedman,² Michael W. W. Adams,³ Gary Siuzdak,⁴ Nitin S. Baliga,⁵ Stephen R. Holbrook,¹ John A. Tainer,^{1,6} and **Steven M. Yannone^{1*}** (SMYannone@lbl.gov)

¹Department of Molecular Biology, Lawrence Berkeley National Laboratory, Berkeley, California;

²Center for Life in Extreme Environments, Portland State University, Portland, Oregon; ³Department of Biochemistry and Molecular Biology, University of Georgia, Athens, Georgia; ⁴Center for Mass Spectrometry, The Scripps Research Institute, La Jolla, California; ⁵Institute for Systems Biology, Seattle, Washington; and ⁶Department of Biochemistry and Molecular Biology, The Scripps Research Institute, La Jolla, California

Project Goals: 1. To develop molecular biology tools to affinity tag *S. solfataricus* genes and reintroduce them into the native organism in a high-throughput manner. 2. To isolate tagged protein complexes and modified proteins from soluble and membrane fractions of *S. solfataricus* extracts. 3. To characterize protein complex components and stoichiometry by 1D/2D gel separation, mass spectrometry, and small angle X-ray scattering (SAXS).

Dynamic protein-protein interactions are fundamental to most biological processes and essential for maintaining homeostasis within all living organisms. Understanding the networks of these protein interactions is of critical importance to understanding the complexities of biological systems. The MAGGIE project was conceived, in part, as a response to the DOE GTL initiative to develop technologies to map the proteomes of model organisms. In this project we are exploiting unique characteristics of members of extremophilic Archaea to identify, isolate, and characterize multi-protein molecular machines. We have teamed expertise in mass spectrometry, systems biology, structural biology, biochemistry, and molecular biology to approach the challenges of mapping relatively simple proteomes. As part of the MAGGIE project, we are developing shuttle vectors for the extremophilic organism *Sulfolobus solfataricus* which has a growth optimum at 80°C and pH 3.0. We are using a naturally occurring viral pathogen of this organism to engineer shuttle vectors designed for recombinant protein tagging and expression in the native *Sulfolobus* background. We are also exploiting the unique characteristics of Archaeal membranes to isolate membrane-protein complexes from native biomass. We will test the idea that the hyperthermophilic nature of *Sulfolobus* will allow U.S. to “thermally trap” protein complexes assembled at 80°C by isolating these complexes at room temperature. Our component is interfacing with other MAGGIE components to characterize isolated proteins and protein complexes using MS/MS and small angle x-ray scattering at the advanced light source at LBNL. Ultimately, we aim to identify metabolic modules suitable to transfer specific metabolic processes between microbes to address specific DOE missions while developing generally applicable molecular and biophysical technologies for GTL.

32 ^{GTL}**Protein Complex Analysis Project (PCAP): Project Overview**

Dwayne Elias,³ Swapnil Chhabra,¹ Jil T. Geller,¹ Hoi-Ying Holman,¹ Dominique Joyner,¹ Jay Keasling,^{1,2} Aindrila Mukhopadhyay,¹ Mary Singer,¹ Tamas Torok,¹ Judy Wall,³ Terry C. Hazen,¹ Gareth Butland,¹ Ming Dong,¹ Steven C. Hall,⁴ Bing K. Jap,¹ Jian Jin,¹ Susan J. Fisher,⁴ Peter J. Walian,¹ H. Ewa Witkowska,⁴ Lee Yang,¹ **Mark D. Biggin**^{1*} (mdbiggin@lbl.gov), Manfred Auer,¹ Agustin Avila-Sakar,¹ Florian Garczarek,¹ Robert M. Glaeser,¹ Jitendra Malik,² Eva Nogales,^{2,4} Hildur Palsdottir,¹ Jonathan P. Remis,¹ Dieter Typke,¹ Kenneth H. Downing,^{1a} Steven S. Andrews,¹ Adam P. Arkin,^{1,2} Steven E. Brenner,^{1,2} Y. Wayne Huang,¹ Janet Jacobsen,² Keith Keller,² Ralph Santos,¹ Max Shatsky,² and John-Marc Chandonia¹

¹Lawrence Berkeley National Laboratory, Berkeley, California; ²University of California, Berkeley, California; ³University of Missouri, Columbia, Missouri; and ⁴University of California, San Francisco, California

Project Goals: The Protein Complex Analysis Project (PCAP) has two major goals: 1. to develop an integrated set of high throughput pipelines to identify and characterize multi-protein complexes in a microbe more swiftly and comprehensively than currently possible and 2. to use these pipelines to elucidate and model the protein interaction networks regulating stress responses in *Desulfovibrio vulgaris* with the aim of understanding how this and similar microbes can be used in bioremediation of metal and radionuclides found in U.S. Department of Energy (DOE) contaminated sites.

The Protein Complex Analysis Project (PCAP) has two major goals: 1. to develop an integrated set of high throughput pipelines to identify and characterize multi-protein complexes in a microbe more swiftly and comprehensively than currently possible and 2. to use these pipelines to elucidate and model the protein interaction networks regulating stress responses in *Desulfovibrio vulgaris* with the aim of understanding how this and similar microbes can be used in bioremediation of metal and radionuclides found in U.S. Department of Energy (DOE) contaminated sites.

PCAP builds on the established research and infrastructure of another Genomics:GTL initiative conducted by the Environmental Stress Pathways Project (ESPP). ESPP has developed *D. vulgaris* as a model for stress responses and has used gene expression profiling to define specific sets of proteins whose expression changes after application of a stressor. Proteins, however, do not act in isolation. They participate in intricate networks of protein / protein interactions that regulate cellular metabolism. To understand and model how these identified genes affect the organism, therefore, it is essential to establish not only the other proteins that they directly contact, but the full repertoire of protein / protein interactions within the cell. In addition, there may well be genes whose activity is changed in response to stress not by regulating their expression level but by altering the protein partners that they bind, by modifying their structures, or by changing their subcellular locations. There may also be differences in the way proteins within individual cells respond to stress that are not apparent in assays that examine the average change in a population of cells. Therefore, we are extending ESPP's analysis to characterize the polypeptide composition of as many multi-protein complexes in the cell as possible and determine their stoichiometries, their quaternary structures, and their locations in planktonic cells and in individual cells within biofilms. PCAP will characterize complexes in wild type cells grown under normal conditions and also examine how these complexes are affected in cells perturbed by stress or by mutation of key stress regulatory genes. These data will all be combined with those of the ongoing work of the ESPP to understand, from a physical-chemi-

cal, control-theoretical, and evolutionary point of view, the role of multi-protein complexes in stress pathways involved in the biogeochemistry of soil microbes under a wide variety of conditions.

Essential to this endeavor is the development of automated high throughput methods that are robust and allow for the comprehensive analysis of many protein complexes. Biochemical purification of endogenous complexes and identification by mass spectrometry is being coupled with *in vitro* and *in vivo* EM molecular imaging methods. Because no single method can isolate all complexes, we are developing two protein purification pipelines, one the current standard Tandem Affinity Purification approach, the other a novel tagless strategy. Specific variants of each of these are being developed for water soluble and membrane proteins. Our Bioinstrumentation group is developing highly parallel micro-scale protein purification and protein sample preparation platforms, and mass spectrometry data analysis is being automated to allow the throughput required. The stoichiometries of the purified complexes are being determined and the quaternary structures of complexes larger than 250 kDa are being solved by single particle EM. We are developing EM tomography approaches to examine whole cells and sectioned, stained material to detect complexes in cells and determine their localization and structures. New image analysis methods will be applied to speed determination of quaternary structures from EM data. Once key components in the interaction network are defined, to test and validate our pathway models, mutant strains not expressing these genes will be assayed for their ability to survive and respond to stress and for their capacity for bioreduction of DOE important metals and radionuclides.

Our progress during the first year of the project includes constructing genetically altered *D. vulgaris* strains and using them to test a range of affinity tags for the purification of protein complexes and EM localization of complexes in cells, developing automated primer design algorithms for high throughput recombinant DNA and strain engineering, establishing cost effective strategies to produce up to 200 L cultures of *D. vulgaris* per week, establishing an optimized four-step tagless fractionation series for the purification of water soluble protein complexes, adopting an efficient PVDF membrane micro titer plate-based methods for mass spectrometry sample preparation, determining the structure of the 1 MDa *D. vulgaris* Pyruvate Ferredoxin Oxidoreductase complex to 17Å resolution by single particle EM, and constructing relational databases for biomass production, genetically manipulated *D. vulgaris* strains, single particle EM, and tagless complex purification. Further details on these and other results are provided in Subproject specific posters.

33 ^{GTL}

Protein Complex Analysis Project (PCAP): Multi-Protein Complex Purification and Identification by Mass Spectrometry

Gareth Butland,¹ Ming Dong,¹ Steven C. Hall,² Bing K. Jap,¹ Jian Jin,¹ Susan J. Fisher,² Peter J. Walian,¹ H. Ewa Witkowska,² Lee Yang,¹ and **Mark D. Biggin**^{1*} (mdbiggin@lbl.gov)

¹Lawrence Berkeley National Laboratory, Berkeley, California and ²University of California, San Francisco, California

Project Goals: This subproject of the Protein Complex Analysis Project (PCAP) is developing several complementary high throughput pipelines to purify protein complexes from *D. vulgaris*, identify their polypeptide constituents by mass spectrometry, and determine their stoichiometries. Our goal is to determine an optimum strategy that may include elements of each purifica-

tion method. These methods will then be used as part of PCAP's effort to model stress responses relevant to the detoxification of metal and radionuclide contaminated sites.

This subproject of the Protein Complex Analysis Project (PCAP) is developing several complementary high throughput pipelines to purify protein complexes from *D. vulgaris*, identify their polypeptide constituents by mass spectrometry, and determine their stoichiometries. Our goal is to determine an optimum strategy that may include elements of each purification method. These methods will then be used as part of PCAP's effort to model stress responses relevant to the detoxification of metal and radionuclide contaminated sites.

Our first purification approach is a novel "tagless" method that fractionates the water soluble protein contents of a bacterium into a large number of fractions, and then identifies the polypeptide composition of a rational sampling of 10,000 – 20,000 of these fractions using MALDI TOF/TOF mass spectrometry. Our second purification approach for water soluble proteins uses and extends the proven Tandem Affinity Purification method (TAP), in which tagged versions of gene products are expressed in vivo and then used to purify the tagged protein together with any other endogenous interacting components. Our third and fourth approaches are specialized variants of the tagless and TAP methods that are being designed to capture membrane protein complexes. A major part of our effort is the design and construction of automated instruments to speed the throughput of protein purification and sample preparation prior to mass spectrometry, and the development of rapid mass spectrometry data analysis algorithms.

Once fully established, we will use our optimized methods to catalog as thoroughly as practicable the repertoire of stable heteromeric complexes in wild type cells grown under normal conditions, as well as identify a number of larger homomeric complexes. We will then examine changes in the composition of protein complexes in cells with perturbed stress response pathways. Response pathways will be perturbed either by growing cells in the presence of stressors, including nitrite, sodium chloride, and oxygen, or by mutating cells to delete a component of a stress response pathway. Purified heteromeric and homomeric complexes larger than 250 kDa are being provided to the EM Subproject to allow their structures to be determined and any stress induced changes in conformation to be detected. All of these data will be correlated by PCAP's Bioinformatics Subproject with computational models of stress response pathways that are currently being established by the Environmental Stress Pathways Project (ESPP).

Our results for the first year of the project are as follows.

Tagless purification of water soluble complexes. We have developed an optimized four-step fractionation scheme for the tagless purification strategy and have used it to identify and purify 15 homomeric and heteromeric water soluble protein complexes from *D. vulgaris*. We have established an efficient, highly reproducible mass spectrometry sample preparation protocol that uses 96-well PVDF multiscreen plates, which will greatly aid high throughput analysis. This sample preparation method is effective for the iTRAQ methodology we have adopted to help quantitate the relative abundances of polypeptides in different chromatographic fractions. Methods for preparing protein samples suitable for single particle EM analysis are being refined, including the use of different crosslinking reagents to stabilize complexes on EM grids. To date, five complexes have been sent to the EM Subproject for structural determination. As a result a 17 Å resolution structure of the 1 MDa Pyruvate Ferredoxin Oxidoreductase complex has been obtained.

Tagless purification of membrane complexes. We have developed an improved strategy to isolate membrane protein complexes that uses a multi-step membrane solubilization approach in which inner and outer membrane proteins are processed sequentially. The choice of detergent was shown

to be critical, especially for the isolation of outer membrane protein complexes. Our methods for preparing samples for mass spectrometry analysis have been improved, particularly in the area of membrane protein native PAGE. Five membrane protein complexes have been identified and several others have been purified, ready for mass spectrometry analysis. Large-scale application of these methods is expected to facilitate the isolation and identification of substantially more complexes over the next project year.

Tandem Affinity Purification of water soluble complexes. We have commenced trials of different TAP tag combinations for protein complex purification from *D. vulgaris*. Initial tests have compared the efficiency and observed non-specific binding properties of Sequential Peptide Affinity (SPA) tag, which is composed of Calmodulin Binding Peptide (CBP) and FLAG affinity purification tags, and the Strep-TEV-FLAG (STF) tag, which is similar to SPA but with CBP replaced by a Streptavidin tag. We have confirmed that both tags can purify proteins synthesized in *D. vulgaris* and are currently testing an expanded set of tagged proteins. Once completed, high-throughput methods currently being developed will be used to construct tagged *D. vulgaris* genes rapidly and efficiently.

Automation of protein complex purification. We have developed a prototype multi-channel, native gel electrophoresis instrument for high resolution protein separation and automated band collection. This instrument can elute a protein band into a 200 µl fraction (about 60% of the band), without noticeable loss of sample. The use of this free-flow electrophoresis apparatus will greatly assist our efforts to achieve high throughput and provide an additional means of obtaining specimens in amounts appropriate for EM studies.

Mass spectrometry. Optimization of MALDI TOF/TOF MS/MS conditions is necessary to maximize the quality and consequently the information content of the data. Resolution of the precursor selection window, number of laser shots, collision energy and collision gas pressure have been evaluated from the point of view of the success rate of protein identification and quality of the iTRAQ-based quantitation. Ultimately our high throughput mass spectrometry workflow will employ highly customized information-dependent selection of precursors for MS/MS. We have begun evaluation of different aspects of iterative MS/MS acquisition routines with the aim of limiting collection of redundant data on proteins already identified and focusing on reliable quantification and identification of less abundant species. The strategy employs collection of MS and very limited MS/MS during the first iteration followed by MS/MS acquisition performed in discrete stages, with each stage building upon a combination of results of current and preceding analyses of adjacent fractions within the same protein complex separation step. We have also are developing algorithms and graphical display tools for identifying protein complexes from mass spectrometry data, including a method for cluster analysis of tagless iTRAQ data to allow for detection of comigrating polypeptides and hence putative protein complexes.

34 ^{GTL}**Protein Complex Analysis Project (PCAP): Imaging Multi-Protein Complexes by Electron Microscopy**

Manfred Auer,¹ Agustin Avila-Sakar,¹ David Ball,¹ Florian Garczarek,¹ Robert M. Glaeser,¹ Jitendra Malik,² Eva Nogales,^{1,2} Hildur Palsdottir,¹ Jonathan Remis,¹ Max Shatsky,² Dieter Typke,¹ and **Kenneth H. Downing**^{1*} (KHDowning@lbl.gov)

¹Lawrence Berkeley National Laboratory, Berkeley, California and ²University of California, Berkeley, California

Project Goals: The broad aim of this Subproject of PCAP is to demonstrate the feasibility of using electron microscopy for high-throughput structural characterization of multi-protein complexes in microbes of interest to DOE. One goal of this work is to characterize the degree of structural homogeneity or diversity of the multi-protein complexes purified by PCAP and to determine the spatial arrangements of individual protein components within the quaternary structure of each such complex. A second goal is to determine the spatial organization and relative locations of large multi-protein complexes within individual, intact microbes. A third goal is to determine whether whole-cell characterization by cryo-tomography can be further supplemented by electron microscopy of cell-envelope fractions and even the whole-cell contents of individual, lysed cells. Finally, plastic-section electron microscopy is used to translate as much as possible of this basic understanding to the more relevant physiological conditions, both stressed and unstressed, of planktonic and biofilm forms of microbes of interest.

The broad aim of this Subproject of PCAP is to demonstrate the feasibility of using electron microscopy for high-throughput structural characterization of multi-protein complexes in microbes of interest to DOE. Protocols and infrastructure are being developed to identify suitable candidates for structural study among the complexes isolated by the other components of the PCAP group, and to determine the spatial arrangements of individual protein components within the quaternary structure of each such complex. At a resolution of ~ 2 nm it is possible to locate the positions of individual proteins within such complexes and to then dock previously-determined atomic models of the identified proteins into the envelope of the density map. At resolutions better than 1 nm it is possible to further characterize conformational changes. We aim to increase the throughput of such structure determinations to the level that quaternary structures and docked atomic models are produced within 48 hours of purification of individual, structurally homogeneous complexes.

A second goal is to determine the spatial organization and relative locations of large multi-protein complexes within individual, intact microbes. It has quite recently been established that cryo-EM tomography can be used to produce clearly distinguishable images of larger multiprotein complexes ($M_r > \sim 750$ k) within suitably thin, intact cells. Since the cells are imaged in a nearly undisturbed condition, it is possible to count the number of such complexes in each cell as well as to characterize their spatial distribution and their association with other components of subcellular structure. Our present aim is to characterize large subcellular structures in *Desulfovibrio vulgaris* to provide a basis for understanding the morphological changes that follow various stresses.

We also employ plastic-section electron microscopy to study both planktonic and biofilm forms of microbes of interest. This approach has the advantage that it lends itself more easily to labeling – and thus localizing – genetically tagged proteins. Sectioning is also the only technique that can provide images of specimens that are too thick to image as whole-mount materials, while still retaining

nanometer resolution. The ultimate goal in using plastic-section microscopy is thus to provide the most complete and accurate information possible about the status of multi-protein complexes, and to do so in a way that can then be used to improve mathematical modeling of cellular responses under the environmental conditions that require bioremediation.

In our initial work on single particle EM structures, we developed a pipeline for characterization of sample homogeneity that calls for an initial evaluation of each such specimen in uranyl acetate, in neutralized phosphotungstic acid, and in ammonium molybdate, in order to minimize misleading characterizations that inevitably occur due to unwanted stain-specimen interactions (e.g. spurious aggregation). One of the more promising specimens that appeared to be suitably homogeneous and well dispersed has been fully analyzed as a demonstration of the procedures and the information provided by the analysis. The protein was identified as pyruvate-ferredoxin oxidoreductase (PFOR), which is present in *D. vulgaris* as an octamer of about 1 MDa molecular weight, while it is found as a dimer in other bacteria, including *Desulfovibrio africanus*. From EM images of protein in negative stain, a three dimensional density map was derived with resolution sufficient to unambiguously dock the x-ray crystal structure of a dimeric form of the enzyme that had previously been obtained from the *D. africanus* protein. The *D. vulgaris* amino acid sequence is found to have one insertion in a surface loop at the interface between subunits that would appear to account for the difference in oligomerization state. Further work is needed to understand the physiological significance of forming the octamer rather than dimer in *D. vulgaris*.

In order to take advantage of the genetic tools that allow tagging of specific proteins for localization by both light and electron microscopy, we are focusing on several fluorescent reagents that can be characterized in the light microscope and then photoconverted to electron dense signals for electron microscopy. This is quite a new endeavor for anaerobic bacteria such as *D. vulgaris*. Our initial focus is on morphology of biofilms in which we see a number of structures that have yet to be characterized in *D. vulgaris*. We grow biofilms of *D. vulgaris* in cellulose dialysis tubing, where the biofilms cover almost the entire interior of the tube. Samples are high pressure frozen and freeze-substituted in order to optimize preservation of structural details. Electron microscopic analysis of biofilm sections reveal loose packing of *D. vulgaris* within the biofilm EPS. Interestingly we found filamentous string-like metal precipitates near the *D. vulgaris*, which may point to structures not unlike the well-characterized *Shewanella* nanowires, which are known to be instrumental in extracellular metal reduction. Variations in the deposition patterns indicate that metal reduction activity varies between neighboring cells in biofilms. Of particular interest are strings of vesicles that appear to be extruded from the cells, similar to structures we have seen also in biofilms of *Shewanella oneidensis* and *Myxococcus xanthus*. We have developed on-grid culturing methods for fast study of such features in cell monolayers grown under various environmental conditions

We have had preliminary success in ReAsH and SNAP-labeling of several strains of *D. vulgaris* in which proteins have been tagged by members of the PCAP Microbiology group. The labeling appears promising as judged by light microscopy, and in-vitro labeling of tagged proteins after cell lysis followed by SDS PAGE suggests specific binding for the SNAP-tag reagent. First attempts at photoconversion of the fluorescence signal are currently underway.

35 ^{GTL}**Protein Complex Analysis Project (PCAP): Microbiology Subproject**

Hoi-Ying Holman,^{1,4} Jay Keasling,^{1,2,4} Aindrila Mukhopadhyay,^{1,4} Swapnil Chhabra,^{1,4} Jil T. Geller,^{1,4} Mary Singer,^{1,4} Dominique Joyner,^{1,4} Tamas Torok,^{1,4} Judy Wall,^{3,4} Dwayne A. Elias,^{3,4} and **Terry C. Hazen**^{1,4*} (tchazen@lbl.gov)

¹Lawrence Berkeley National Laboratory, Berkeley, California; ²University of California, Berkeley, California; ³University of Missouri, Columbia, Missouri; and ⁴Virtual Institute for Microbial Stress and Survival, Berkeley, California <http://vimss.lbl.gov>

Project Goals: The Microbiology Subproject of the Protein Complex Analysis Project (PCAP) provides the relevant field experience to suggest the best direction for fundamental, but DOE relevant, research as it relates to bioremediation and natural attenuation of metals and radionuclides at DOE contaminated sites. We are building on techniques and facilities established by the Environmental Stress Pathways Project (ESPP) for isolating, culturing, and characterizing *Desulfovibrio vulgaris*. The appropriate stressors for study will be identified and, using stress response pathway models from ESPP, the relevance and feasibility for high throughput protein complex analyses will be assessed. Two types of genetically engineered strain are being constructed: strains expressing affinity tagged proteins and knock out mutation strains that eliminate expression of a specific gene. High throughput phenotyping of these engineered strains will then be used to determine if any show phenotypic changes. Finally, we are producing large quantities of cells under different conditions and harvesting techniques for optimal protein complex analyses.

The Microbiology Subproject of PCAP provides the relevant field experience to suggest the best direction for fundamental, but DOE relevant research as it relates to bioremediation and natural attenuation of metals and radionuclides at DOE contaminated sites. This project builds on techniques and facilities established by the Virtual Institute for Microbial Stress and Survival (VIMSS) for isolating, culturing, and characterizing *Desulfovibrio vulgaris*. The appropriate stressors for study have been identified and, using stress response pathway models from VIMSS, the relevance and feasibility for high throughput protein complex analyses is being assessed. We also produce all of the genetically engineered strains for PCAP. Two types of strain are being constructed: strains expressing affinity tagged proteins and knock out mutation strains that eliminate expression of a specific gene. We anticipated producing over 300 strains expressing affinity tagged proteins every year for complex isolation and EM labeling experiments by the other Subprojects. A much smaller number of knockout mutation strains are being produced to determine the effect of eliminating expression of components of putative stress response protein complexes. Both types of engineered strains are being generated using a two-step procedure that first integrates and then cures much of the recombinant DNA from the endogenous chromosomal location of the target gene. We are developing new counter selective markers for *D. vulgaris*. This procedure will 1) allow multiple mutations to be introduced sequentially, 2) facilitate the construction of in-frame deletions, and 3) prevent polarity in operons. The Microbiology Subproject provides high throughput phenotyping of all engineered strains to determine if any show phenotypic changes. We also determine if the tagged proteins remain functional and that they do not significantly affect cell growth or behavior. The knockout mutations are tested in a comprehensive set of conditions to determine their ability to respond to stress. High throughput optimization of culturing and harvesting of wild type cells and all engineered strains are used to determine the optimal time points, best culture techniques, and best techniques for harvesting cultures using real-time analyses with synchrotron FTIR spectromicroscopy, and other methods. Finally, we are producing large quantities of cells under different conditions and harvesting techniques for opti-

mal protein complex analyses. To insure the quality and reproducibility of all the biomass for protein complex analyses we use extreme levels of QA/QC on all biomass production. We expect to do as many as 10,000 growth curves and 300 phenotype microarrays annually and be producing biomass for 500-1000 strains per year by end of the project. Each biomass production for each strain and each environmental condition will require anywhere from 0.1 – 400 L of culture, and we expect more than 4,000 liters of culture will be prepared and harvested every year. The Microbiology Subproject is optimizing phenotyping and biomass production to enable the other Subprojects to complete the protein complex analyses at the highest throughput possible. Once the role of protein complexes has been established in the stress response pathway, we will verify the effect that the stress response has on reduction of metals and radionuclides relevant to DOE.

During the last year, the Microbiology Subproject has supplied more than 30 sets of *D. vulgaris* cultures on biofilms for EM analysis, more than twenty 1-5 L cultures of biomass for water-soluble protein complex purification studies, and two 100 L and one 200 L cultures for membrane protein complex purification. We have designed and implemented a continuous culture system that enables U.S. to produce more than 300 L of anaerobic mid log phase *D. vulgaris* in as little as 5 days, including harvesting and QA techniques that maximize reproducibility of all biomass produced. The goal of incorporating different affinity or tandem affinity (TAP) protein tags into three genes to determine the best tag for use in the PCAP project will be complete by the end of 2006. These include the Strep-tag® (IBA) for streptavidin-binding, the SPA-tag (a.k.a. CTF) that consists of a calmodulin binding motif, tobacco etch virus protease (TEV) and 3X FLAG affinity, as well as a combination of these that replaces the calmodulin binding with the Strep-tag® resulting in STF. The three genes to test are the dissimilatory sulfite reductase subunit C, pyruvate ferredoxin oxidoreductase subunit B and ATP synthase subunit C. Additionally, several other gene targets have been identified through close collaboration with the VIMSS/ESPP group at LBNL and are currently being tagged. To determine localization of a given gene product in the cell, we have utilized both the tetracycline and SNAP-tag™ (Covalys) in cooperation with the EM group of the PCAP project. Currently the total number of genes tagged with CTF are 3, with STF 6, with strep 20, with tetracycline 8 and with SNAP 3. We are currently attempting to construct an ordered library for tagging in DvH. By doing so, this will allow for a relatively small number of *E. coli* clones to carry all of the genes for DvH, thereby reducing the overall workload and paving the way for higher throughput tagging. We also compared several cloning strategies for producing tagged constructs for the plasmid insertion strategy in *D. vulgaris*. The two-step TOPO-GATEWAY strategy (Invitrogen) was identified as the most economical commercially available conventional-cloning strategy amongst these. We also developed a workflow for high throughput production of tagged strains of *D. vulgaris*. This workflow was based on the TOPO-GATEWAY strategy in combination with current technology for transformation of *D. vulgaris*. We constructed custom destination vectors carrying the tags SPA and STF (to realize the GATEWAY step)—these are not available commercially. We tested the workflow through all the steps for a set of 10 randomly chosen genes from the *D. vulgaris* genome. Based on generated sequences, five of these were successfully tagged with the SPA tag. We also tested a commercially available 96-well electroporation device (BTX) for high throughput transformation of *D. vulgaris*. This system was found unsuitable. We are currently developing a custom solution for this. We also collaborated with the Subgroup D (Computational Core) for the development of automated algorithms for: 1) Primer identifications based on gene locations within operons for PCR amplifications in 96-well format, and 2) Sequence alignments for identifying errors in the amplifications or cloning steps in the workflow. All of the tagged strains constructed this year have been characterized using phenotypic microarrays (PM), and the *D. vulgaris* megaplasmid minus strain (MP(-)) being used in the electroporation studies is being aggressively characterized for all differences including stress responses by the ESPP project.

36 ^{GTL}**Protein Complex Analysis Project (PCAP): High Throughput Strategies for Tagged-Strain Generation in *Desulfovibrio vulgaris***

Swapnil Chhabra^{1*} (SRChhabra@lbl.gov), Gareth Butland,¹ Dwayne Elias,² Aindrila Mukhopadhyay,¹ John-Marc Chandonia,¹ Jay Keasling,^{1,3} and **Judy Wall**²

¹Lawrence Berkeley National Laboratory, Berkeley, California; ²University of Missouri, Columbia, Missouri; and ³University of California, Berkeley, California

Project Goals: As part of the microbiology core of the Protein Complex Analysis Project (PCAP) our goal is to develop a technological platform for creating a library of *D. vulgaris* mutant strains expressing tagged proteins at high throughput. Based on the workflow designed around the TOPO-GATEWAY strategy, we will produce a hundred constructs carrying the STF tag which will be transformed in *D. vulgaris* to create a tagged strain library. We are also exploring an alternative high throughput strategy using an ordered library of *D. vulgaris*.

Most cellular processes are mediated by a host of different proteins interacting with each other in the form of complexes. As a follow-up to the functional genomics analyses of stress response pathways in *Desulfovibrio vulgaris*, are studying the role of protein complexes in this sulfate reducing bacterium which has been found to exist in several DOE waste sites. As part of the microbiology core of the Protein Complex Analysis Project (PCAP) our current goal is to develop a technological platform for creating a library of *D. vulgaris* mutant strains expressing tagged proteins. We are currently exploring two approaches to achieve this goal. The first strategy involves the use of plasmid constructs carrying single target genes using the two-step TOPO-GATEWAY cloning approach (Invitrogen). The first step in the strategy involves generation of an entry vector carrying the gene of interest (GOI) via TOPO cloning. The second step involves transfer of the GOI from the entry vector to a suitable destination vector (carrying the tag of choice) through an in-vitro recombination reaction. The second strategy involves the use of an ordered library of *D. vulgaris* modified using a lambda-red phage system. Library constructs are modified, in a strain of *E. coli* expressing the lambda-red recombination system, using linear PCR products specifically engineered to recombine into the 3' end of the gene of interest. These PCR products, when inserted into the gene of interest modify the coding sequence of the gene to encode a C-terminal fusion protein bearing the tag of choice. Constructs from both strategies are transferred to *D. vulgaris* via electroporation and replacement of the wild type copy of the gene of interest with the tagged version is selected. These approaches would enable U.S. to rapidly and efficiently modify the *D. vulgaris* genome and express tagged proteins at their native levels.

Protein Interaction Reporters and Outer Membrane Cytochrome C

James E. Bruce^{1*} (james_bruce@wsu.edu), Haizhen Zhang,¹ Natalia Zakharova,¹ Xiaoting Tang,¹ Gerhard R. Muske,¹ Liang Shi,² James K. Fredrickson,² Nikola Tolic,² and Gordon A. Anderson²

¹Department of Chemistry, Washington State University, Pullman, Washington <http://www.wsu.edu/ proteomics> and ²Environmental Molecular Science Laboratory, Pacific Northwest National Laboratory, Richland, Washington <http://www.emsl.pnl.gov>

Project Goals: The overall aim of this project is to develop the PIR approach to identify protein interactions in cells and apply this new methodology to the study of systems relevant to the DOE mission.

We have developed a unique chemical cross-linking system that employs novel compounds that we call “Protein Interaction Reporters” or PIRs that can help identify interactions among proteins in complex biological systems. The incorporation of mass spectrometry-cleavable bonds in the PIR structure allows release of labeled, intact peptides that can subsequently be analyzed and identified with high mass measurement accuracy or tandem MS methods. This approach enables a new method of analysis of cross-linked proteins from complex mixtures since protein identification is established via released peptide measurements, while interaction information is established through PIR-labeled peptide analysis. Since interactions among proteins in cells are critical determinants of overall function, the ability to identify and measure protein interactions in a cellular system with PIR technology can significantly improve molecular-level comprehension of biological function. The overall aim of this project is to develop the PIR approach to identify protein interactions in cells and apply this new methodology to the study of systems relevant to the DOE mission. One area where this approach will have positive impact is in the study of electron transfer mechanisms in bacteria such as *Shewanella oneidensis*. For example, a central question relevant to many biological systems with potential for bioremediation or bio-energy production revolves around the protein interaction pathways that facilitate novel mechanisms of electron transport. Improved understanding of these pathways can conceivably guide bioengineering efforts to result in improved electron transport properties.

We have previously applied the PIR strategy to map interactions in a model noncovalent complex and illustrate feasibility¹. More recently, we have applied PIR technology to the microbial system, *Shewanella oneidensis* MR-1, and identified more than 380 proteins that are labeled during on-cell PIR reactions². We have also developed additional PIR structures that provide increased chemical diversity to further improve the number and type of proteins that can be studied with the PIR approach³. This report will present these recent developments and the application of the PIR strategy with efforts focused on the identification of interactions of proteins known to be critical to electron transport, such as outer membrane cytochrome (OMC) proteins. This report will also highlight our efforts to study OMC proteins and their interactions both *in vitro* and in cells using the PIR strategy combined with immunoaffinity methods.

This research was supported by the Office of Science (BER), U.S. Department of Energy, Grant No. DE-FG02-04ER63924.

References

1. Tang, X., Munske, G.R., Siems, W.F. & Bruce, J.E. Mass spectrometry identifiable cross-linking strategy for studying protein-protein interactions. *Anal Chem* **77**, 311-318 (2005).

2. Tang, X. Yi, W., Munske, G.R. Adhikari, D.P. Zakharova, N.L. Bruce, J.E. Profiling the membrane proteome of *Shewanella oneidensis* MR-1 with new affinity labeling probes. *J. Proteome Research*, in press.
3. Chowdhury, S.M. Munske, G.R. Tang, X. Bruce, J.E. Collisionally Activated Dissociation and Electron Capture Dissociation of Several Mass Spectrometry-Identifiable Chemical Cross-Linkers. *Anal. Chem.* in press.

38 [—]_{GTL}

Technologies for Comprehensive Protein Production

Sarah Giuliani, Elizabeth Landorf, Terese Peppler, and **Frank Collart*** (fcollart@anl.gov)

Argonne National Laboratory, Argonne, Illinois

Progress in genome sequencing has accentuated the importance of high throughput proteomic strategies for identification of cellular function. Many proteomic technologies such as functional screens, structure determination or interaction mapping use purified proteins as a starting point. Although, recent advances in expression technology have significantly increased our capability for the expression of microbial proteins, a significant fraction of the most interesting proteins encoded by the genome still cannot be expressed in an experimentally usable form. We have developed technologies to optimize the expression of proteins and protein domains from prokaryotic and eukaryotic organisms. Purity will be assessed by SDS-PAGE stained with Coomassie Brilliant Blue. These proteins are a valuable resource for characterization studies and for structural and functional studies.

One application of this resource is the development of new approaches for in vitro production, validation, and characterization of protein complexes. Identification and characterization of protein complexes is an important component of several GTL science programs. The complexity of bacterial proteome and our limited knowledge of the number and nature of interacting proteins indicate multiple strategies must be applied to obtain a complete understanding of the number and function of cellular interacting proteins. In a preliminary study, we identified a preliminary set of 50 putative interacting proteins in *S. oneidensis* based on homology to known interacting proteins from other bacteria. These putative interacting protein pairs were validated for complex formation using an in vitro interaction assay in a manual format and via high density protein arrays for identifying interactions on protein biochips. The results indicate that many of the interactions found in *E. coli* could be rapidly validated for other organism using a high throughput approach. This set of putative interacting proteins is also being used to development of more efficient approaches for co-expression of proteins. Current expression systems sometimes produce protein subunits or interacting proteins in an insoluble form. We intend use interaction pairs from the primary screen with an insoluble or low solubility partner will be screened for improvement in production of soluble protein complexes using two co-expression strategies.

39 ^{GTL}

Next-Generation Cell-Permeable Multiuse Affinity Probes (MAPs) and Cognate Tags: Applications to Bioenergy and Metabolic Engineering

M. Uljana Mayer, Baowei Chen, Haishi Cao, Ting Wang, Ping Yan, Yijia Xiong, Liang Shi, and **Thomas C. Squier*** (thomas.squier@pnl.gov)

Pacific Northwest National Laboratory, Richland, Washington

Project Goals: Our long-term goal is to develop the necessary reagents and technology for the rapid identification of a substantial fraction of the multiprotein complexes in an organism, thereby permitting a systems level understanding of signaling complexes that allow microorganisms to adapt to diverse environmental conditions. Critical to this goal is the development of new affinity reagents and protein tags that permit the rapid identification and validation of protein complexes. To accomplish these goals, we propose to develop protein tags and associated multiuse affinity probes, which promise both the ability to isolate and characterize protein complexes as well as the examination of protein-protein interactions in living microbes. We propose to extend the design of affinity reagents to include cell permeable and fluorescent photo-crosslinking reagents that permit efficient stabilization and visualization of bacterial protein complexes. Crosslinking reagents will be designed that i) react with known chemistries targeting specific protein side-chains with high yield and ii) provide cleavage sites to facilitate mass spectrometric identification of crosslinked peptides. Initial efforts will focus these technologies to identify protein complexes in *S. oneidensis* MR-1, an organism of considerable interest to DOE and the subject of multiple projects in the Genomics:GTL program (<http://genomicsgtl.energy.gov/>). Utilization of novel affinity reagents that become fluorescent upon binding to engineered tags will permit quantitation of expressed proteins and purification and stabilization of protein complexes. Ultimately high-throughput data collection of time-dependent changes in protein complex formation in response to environmental conditions will be available, thus permitting a systems level understanding of how microbes adapt to environmental change.

Protein-protein interactions are the foundation of the metabolic and regulatory pathways in all organisms. Before organisms can be harnessed in bioremediation and bioenergy uses, robust and scalable methodologies are needed that permit the facile identification of these pathways. To this purpose, we have developed a multi-use affinity probe (MAP) technology platform in which engineering a single encoded peptide tag onto a protein permits the identification and validation of protein complexes *in vitro* and *in vivo* in ways amenable to high-throughput scale-up through the addition of robotics. This platform overcomes prior limitations regarding the need to reengineer and perturb biomolecular systems for proteomic and imaging applications. In previous years, we have matured our technology for (i) protein complex isolation and (ii) *in vitro* validation, as well as showing how photoactivatable crosslinking moieties appended to MAPs provide a means to stabilize low-affinity binding proteins prior to cell lysis, providing the first robust means to mediate *in-vivo* cross-linking. In addition, following covalent modification fluorophore-assisted light inactivation (FALI) permits the selective knock-out of protein activities to aid in both the identification of their functions as well as the modulation of metabolic networks (1).

During the last year, we have extended the toolkit by adding new, bright photostable fluorescent probes based on a cyanine scaffold which are targeted to unique peptide tags and can thus be used simultaneously with first generation probes for parallel processing of tagged proteins to i) isolate and

identify the composition and size of multiple protein complexes and ii) image protein locations and associations in cells using multicolor and single molecule measurements (2–6). Further, the MAP toolkit was used to discover new aspects of bacterial transcription regulatory machinery and to isolate and identify functions of outer membrane cytochromes associated with electron transfer to iron oxides (e.g., hematite) from *Shewanella oneidensis*, which has important implications with respect to both bioremediation and bioenergy (4, 7–10).

MAPs permit the isolation of intact protein complexes, whose release from affinity matrices by mild reducing conditions allows complementary structural and functional measurements, while retaining weak binding interactions (9, 10). Among the important applications of this technology, extensive binding interactions between protein machinery (e.g., RNA polymerase) and regulatory proteins have been identified that are altered in response to environmental changes under controlled growth conditions in chemostats, and have important implications with respect to metabolic engineering principles (9, 10). We find that RNA polymerase from *Shewanella oneidensis* exists as a large supramolecular complex with an apparent mass in excess of 1.4 MDa, whose protein composition substantially changes in response to growth conditions. In comparison to suboxic conditions, a larger number of binding partners associate with RNA polymerase under aerobic conditions, where cellular growth rates are limited by the rates of ribosome biosynthesis (11, 12). In addition to known regulators of RNA polymerase function, binding partners include a surprisingly large number of metabolic enzymes associated with ATP synthesis, nucleotide metabolism, and the biosynthesis of stable RNA (i.e., tRNA and rRNA). Our identification of cytosolic subunits of membrane proteins with the RNA polymerase complex is consistent with recent structural data demonstrating the membrane localization of RNA polymerase in rapidly growing cells (13). In contrast, under suboxic conditions we observe a reduced number of protein associations, which is consistent with the observed disruption of RNA polymerase in condensed structures near the membrane under suboptimal growth conditions (13). In conclusion, these measurements demonstrate an unexpected functional linkage between RNA polymerase and enzymes associated with tRNA processing, nucleotide metabolism, and energy biosynthesis, which we propose to be necessary for optimal transcriptional rates and dependent on growth conditions.

Likewise, we have used this technology to isolate and identify the functions of outer member cytochromes OmcA and MtrC associated with electron transfer to iron oxides (e.g., hematite) from (4, 7, 8). Complementary structural and functional measurements are possible through the ability of the identified tags to promote selective binding to metal surfaces (14). OmcA was shown to directly associate with added hematite, as evidenced by the co-sedimentation of approximately 40% of the OmcA in solution with hematite particles upon centrifugation, permitting a determination of the binding affinity through the measurement of the concentrations of OmcA bound to hematite relative to that in water (i.e., the partition coefficient). We find that approximately 0.2 mg OmcA binds per mg of hematite (i.e., 2.5 nmol OmcA per mg hematite). In association with hematite, OmcA is catalytically active: oxidation of protein hemes, as measured from time-dependent changes in the α -Soret absorption peak at 552 nm, directly tracks with protein binding to hematite under anoxic conditions with a maximal activity of about 60 nmol mg⁻¹ OmcA min⁻¹. Since OmcA can be directly reduced by NADH and other metabolic cofactors (7), the high-affinity interaction between OmcA and hematite provides a means to couple the generation of reducing power to an electrode surface. In summary, we have shown that purified OmcA binds and densely covers the surface of hematite, and reduces Fe(III) with a maximal velocity of approximately 60 nmol / mg min, which corresponds to an electron flux of about 10¹³ electrons /cm²/s that approaches observed fluxes in the most efficient bioreactors (15, 16).

References

1. Yan, P., Xiong, Y., Chen, B., Negash, S., Squier, T. C., and Mayer, M. U. (2006) Fluorophore-assisted light inactivation of calmodulin involves singlet-oxygen mediated cross-linking and methionine oxidation. *Biochemistry* 45, 4736-48.
2. Cao, H., Xiong, Y., Wang, T., Chen, B., Squier, T. C., and Mayer, M. U. (2007) A Cy3-based biarsenical fluorescent probe with a unique peptide binding motif., *J. Am. Chem. Soc.*, *submitted*.
3. Stenoien, D., Knyushko, T., Londono, M., Opresko, L., Mayer, M. U., Squier, T. C., and Bigelow, D. J. (2007) Cellular trafficking of phospholamban and formation of functional sarcoplasmic reticulum during myocyte differentiation. *Am. J. Physiol.*, *submitted*.
4. Xiong, Y., Shi, L., Chen, B., Mayer, M. U., Lower, B. H., Londer, Y., Bose, S., Hochella, M. F., Fredrickson, J. K., and Squier, T. C. (2006) High-affinity binding and direct electron transfer to solid metals by the *Shewanella oneidensis* MR-1 outer membrane c-type cytochrome OmcA. *J Am Chem Soc* 128, 13978-9.
5. Chen, B., Mayer, M. U., Cao, H., Yan, P., Mahaney, J. E., and Squier, T. C. (2007) Selective labeling of cytosolic and membrane proteins using cell-permeable biarsenical probes ReAsH-EDT2 and FIAsh-EDT2 *Anal. Biochem.*, *submitted*.
6. Chen, B., Mayer, M. U., and Squier, T. C. (2007) Identification of an orthogonal peptide binding motif for biarsenical dyes. . *Bioconjugate Chemistry*, *submitted*.
7. Shi, L., Chen, B., Wang, Z., Elias, D. A., Mayer, M. U., Gorby, Y. A., Ni, S., Lower, B. H., Kennedy, D. W., Wunschel, D. S., Mottaz, H. M., Marshall, M. J., Hill, E. A., Beliaev, A. S., Zachara, J. M., Fredrickson, J. K., and Squier, T. C. (2006) Isolation of a high-affinity functional protein complex between OmcA and MtrC: Two outer membrane decaheme c-type cytochromes of *Shewanella oneidensis* MR-1. *J Bacteriol* 188, 4705-14.
8. Shi, L., Lin, J. T., Markillie, L. M., Squier, T. C., and Hooker, B. S. (2005) Overexpression of multi-heme C-type cytochromes. *Biotechniques* 38, 297-9.
9. Mayer, M. U., Shi, L., and Squier, T. C. (2005) One-step, non-denaturing isolation of an RNA polymerase enzyme complex using an improved multi-use affinity probe resin. *Mol Biosyst* 1, 53-6.
10. Verma, S., Xiong, Y., Mayer, M. U., and Squier, T. C. (2007) Remodeling of bacterial RNA polymerase complex in response to environmental conditions *Biochemistry*, *submitted*.
11. Cabrera, J. E., and Jin, D. J. (2003) The distribution of RNA polymerase in *Escherichia coli* is dynamic and sensitive to environmental cues. *Mol Microbiol* 50, 1493-505.
12. Liu, M., Durfee, T., Cabrera, J. E., Zhao, K., Jin, D. J., and Blattner, F. R. (2005) Global transcriptional programs reveal a carbon source foraging strategy by *Escherichia coli*. *J Biol Chem* 280, 15921-7.
13. Jin, D. J., and Cabrera, J. E. (2006) Coupling the distribution of RNA polymerase to global gene regulation and the dynamic structure of the bacterial nucleoid in *Escherichia coli*. *J Struct Biol*, doi:10.1016/j.jsb.2006.07.005.
14. Wigginton, N. S., Rosso, K. M., Lower, B. H., Shi, L., and Hochella, M. F. (2007) Electron tunneling properties of outer membrane decaheme cytochromes from *Shewanella oneidensis*., *Geochimica et Cosmochimica Acta*, *in press*.
15. Ringeisen, B. R., Henderson, E., Wu, P. K., Pietron, J., Ray, R., Little, B., Biffinger, J. C., and Jones-Meehan, J. M. (2006) High power density from a miniature microbial fuel cell using *Shewanella oneidensis* DSP10. *Environ Sci Technol* 40, 2629-34.
16. Viamajala, S., Peyton, B. M., Apel, W. A., and Petersen, J. N. (2002) Chromate/nitrite interactions in *Shewanella oneidensis* MR-1: evidence for multiple hexavalent chromium [Cr(VI)] reduction mechanisms dependent on physiological growth conditions. *Biotechnol Bioeng* 78, 770-8.

40 ^{GT}L

Advanced Data Analysis Pipeline for Determination of Protein Complexes and Interaction Networks at the Genomics:GTL Center for Molecular and Cellular Systems

Kevin K. Anderson,² Deanna L. Auberry,² William R. Cannon^{2*} (William.Cannon@pnl.gov), Don S. Daly,² Brian S. Hooker,² Gregory B. Hurst,¹ Jason E. McDermott,² W. Hayes McDonald^{*1} (McDonaldWH@ornl.gov), Dale A. Pelletier,¹ Denise D. Schmoyer^{1*} (SchmoyerDD@ornl.gov), Julia L. Sharp,³ Mudita Singhal^{2*} (Mudita.Singhal@pnl.gov), Ronald C. Taylor^{2*} (Ronald.Taylor@pnl.gov), **Michelle V. Buchanan**¹ (BuchananMV@ornl.gov)

¹Oak Ridge National Laboratory, Oak Ridge, Tennessee; ²Pacific Northwest National Laboratory, Richland, Washington; and ³Montana State University, Bozeman, Montana

Project Goals: The Center for Molecular and Cellular Systems (CMCS) focuses primarily on the objectives outlined in Goal 1 of the DOE Genomics:GTL program. The core of the CMCS is a high throughput pipeline for the identification of protein complexes, currently focusing on *Rhodopseudomonas palustris*. The pipeline employs affinity isolation coupled with mass spectrometry to identify protein interactions. Computational tools are used to assess the significance of identified interactions. A dynamic research program supports the goals of the CMCS by focusing on the development and implementation of improved capabilities for complex isolation, molecular level identification of the complexes, and critical bioinformatics and computing capabilities. These efforts are focused on constructing a knowledge base that can provide insight into the relationship between the complement of protein complexes in these microbes and their biological function.

The Genomics:GTL Center for Molecular and Cellular Systems (CMCS) is a DOE Center whose mission is to determine protein complexes and interaction networks from microbial systems. Currently the center is focusing on protein interactions involved in nitrogen fixation and metabolism in *Rhodopseudomonas palustris*. The center uses an affinity purification approach (refer to poster **Global survey of protein-protein interactions in *Rhodopseudomonas palustris***) to identify protein interactions in a robust, high-throughput manner. In this process, bait proteins along with co-purifying prey proteins are extracted from the cellular milieu. The resulting protein mixture is analyzed by HPLC coupled with tandem mass spectrometry.

The pipeline for data analysis begins with a laboratory information management system (LIMS) to capture links to the MS/MS data, peptides/proteins identified from those data, and descriptions regarding biological and assay conditions (metadata). The LIMS is the central data repository for all information related to processing and analysis of CMCS samples. It maintains a detailed history for each sample by capturing processing parameters, protocols, stocks, QA/QC tests and analytical results for the complete life cycle of the sample. Project and study data are also maintained to define each sample in the context of the research tasks it supports.

The resulting lists of potentially interacting prey proteins identified from MS/MS are statistically analyzed within a software environment (*Sebini*) specifically designed for working with biological networks. The prey protein lists are cross-tabulated by bait protein to form a prey-by-bait frequency matrix. The frequency pattern across a given row (prey) shows the associations between a prey and the baits. Interpretation of this pattern depends on the selected baits. Pattern uniformity is tested with a binomial-based likelihood-ratio test. Test significance is assessed by Monte Carlo simulation

where the false discovery rate is controlled. Prey protein candidates are assigned to "specific" and "non-specific" classes based on the likelihood-ratio test. Bayes estimates of the confidence of the inferred associations are estimated for each bait-prey pair. Modeling assumptions are investigated and conservative parameter estimates are made using Monte Carlo simulations.

The resulting protein networks are captured in a database within the software environment where information on the nodes (proteins) and edges (interactions) is linked to external and internal bioinformatics data, such as information on interologs derived from the *Bioverse* system, which provides additional information on protein interactions derived from orthologous proteins in other model systems. The joint analysis of experimental data and multiple sources of bioinformatically-derived information is accomplished through collective analysis of biological interaction networks (*Cabin*), a plug-in for the Cytoscape program. Protein interaction networks along with relevant data captured at multiple stages of the data analysis pipeline will be available for download at the project website (Refer to poster **The Microbial Protein-Protein Interaction Database-MiPPI**). A demonstration will be held at the workshop.

41 ^{GTL}

Global Survey of Protein-Protein Interactions in *Rhodopseudomonas palustris*

Dale A. Pelletier^{1*} (pelletierda@ornl.gov), Gregory B. Hurst,¹ Linda J. Foote,¹ Trish K. Lankford,¹ Catherine K. McKeown,¹ Tse-Yuan S. Lu,¹ Elizabeth T. Owens,¹ Denise D. Schmoyer,¹ Manesh B. Shah,¹ Jennifer L. Morrell-Falvey,¹ Brian S. Hooker,² Stephen J. Kennel,¹ W. Hayes McDonald,¹ Mitchel J. Doktycz,¹ Deanna L. Auberry,² William R. Cannon,² Kenneth J. Auberry,² H. Steven Wiley,² and **Michelle V. Buchanan**¹

¹Oak Ridge National Laboratory, Oak Ridge, Tennessee and ²Pacific Northwest National Laboratory, Richland, Washington

Project Goals: The Center for Molecular and Cellular Systems (CMCS) focuses primarily on the objectives outlined in Goal 1 of the DOE Genomics:GTL program. The core of the CMCS is a high throughput pipeline for the identification of protein complexes, currently focusing on *Rhodopseudomonas palustris*. The pipeline employs affinity isolation coupled with mass spectrometry to identify protein interactions. Computational tools are used to assess the significance of identified interactions. A dynamic research program supports the goals of the CMCS by focusing on the development and implementation of improved capabilities for complex isolation, molecular level identification of the complexes, and critical bioinformatics and computing capabilities. These efforts are focused on constructing a knowledge base that can provide insight into the relationship between the complement of protein complexes in these microbes and their biological function.

The goal of the Center for Molecular and Cellular Systems (CMCS) is to identify protein-protein interaction networks, which form the molecular basis of biological function, in environmentally relevant bacterial species in support of the Genomics:GTL program. *Rhodopseudomonas palustris* is a metabolically diverse anoxygenic phototrophic bacterium that is emerging as a model system for nitrogenase-mediated hydrogen production. This process requires several metabolic and regulatory networks to be integrated within the cell, including nitrogen metabolism, photosynthesis and carbon metabolism. To better understand the interactions among these processes, we have begun mapping

protein-protein interactions of photoheterotrophically grown *R. palustris*. Toward this goal, we have developed and implemented a methodology for systematically identifying the proteins that interact with an affinity-tagged “bait” protein expressed from a plasmid introduced into *R. palustris*. The steps in this methodology include target or “bait” selection, primer design, PCR amplification, cloning, transformation, batch culture, lysis, affinity isolation, protein identification, and statistical filtering. Here we will present results on interactions identified by this approach.

To date, we have successfully cloned ~1000 *R. palustris* open reading frames, purified over 250 different affinity-tagged gene products, and identified their protein interaction partners from cultures of *R. palustris* grown under anaerobic photoheterotrophic growth conditions, in the presence or absence of fixed nitrogen. Interactors identified by this approach include homologues of a number of well-characterized protein complexes involved in known metabolic networks, such as nitrogen metabolism (Mo-nitrogenase, FixABCX with a possible role in electron transfer to nitrogenase, GlnK2-AmtB2, GlnK2-GlnB), carbon metabolism (2-oxoglutarate dehydrogenase, succinate dehydrogenase, succinyl-CoA synthetase, tryptophan synthase), transcription (DNA-directed RNA polymerase), chaperones (GroES-EL, DnaK-GrpE), and energy generation (F1F0 ATPase, subunits of NADH dehydrogenase). Novel putative interactions were also identified, including interactions among four anaerobically induced proteins encoded by RPA2334, RPA2335, RPA2336 and RPA2338; interactions between a conserved unknown RPA3193 and a putative acetyltransferase RPA3194; and interactions among conserved unknowns RPA1244, RPA1243 and RPA1246.

We are applying other approaches, including experimental, literature-based and bioinformatics predictions to verify these high-throughput interaction data (refer to poster **Advanced Data Analysis Pipeline for Determination of Protein Complexes and Interaction Networks at the Genomics: GTL Center for Molecular and Cellular Systems**). A web resource for these data will be publicly available in February 2007 (refer to poster **The Microbial Protein-Protein Interaction Database-MiPPI**; www.ornl.gov/sci/GenomestoLife/). These results demonstrate the utility of data emerging from the CMCS for confirming known interactions, as well as for generating hypotheses about potentially novel protein-protein interactions. The identification of protein interactions will aid in elucidation of biological interaction networks and possibly in predicting protein function.

42 ^{GTL}

The Microbial Protein-Protein Interaction Database (MiPPI)

Denise D. Schmoyer^{1*} (schmoyerdd@ornl.gov), Sheryl A. Martin,¹ Gregory B. Hurst,¹ Manesh B. Shah,¹ Dale A. Pelletier,¹ W. Hayes McDonald,¹ William R. Cannon,² Deanna L. Auberry,² and **Michelle V. Buchanan¹**

¹Oak Ridge National Laboratory, Oak Ridge, Tennessee and ²Pacific Northwest National Laboratory, Richland, Washington

Project Goals: The Center for Molecular and Cellular Systems (CMCS) focuses primarily on the objectives outlined in Goal 1 of the DOE Genomics:GTL program. The core of the CMCS is a high throughput pipeline for the identification of protein complexes, currently focusing on *Rhodospseudomonas palustris*. The pipeline employs affinity isolation coupled with mass spectrometry to identify protein interactions. Computational tools are used to assess the significance of identified interactions. A dynamic research program supports the goals of the CMCS by focusing on the development and implementation of improved capabilities for complex isolation, molecular

level identification of the complexes, and critical bioinformatics and computing capabilities. These efforts are focused on constructing a knowledge base that can provide insight into the relationship between the complement of protein complexes in these microbes and their biological function.

The Microbial Protein-Protein Interaction Database (MiPPI) is a publicly accessible database of microbial protein-protein interactions experimentally detected at the Genomics:GTL Center for Molecular and Cellular Systems (CMCS). The primary experimental method used at the CMCS is affinity-based isolation combined with mass spectrometry (refer to poster **Global Survey of Protein-Protein Interactions in *Rhodopseudomonas palustris***). As of December 2006 we have performed over 500 endogenous affinity-tagged experiments which represent over 300 different bait proteins in *Rhodopseudomonas palustris* and *Shewanella oneidensis*. Our goal is to provide the highest quality protein interaction data to the biological community for the identification of cellular networks and ultimately biological function.

MiPPI stores the results of mass spectrometric protein identifications as DTASelect output files, as well as statistical evaluation of protein-protein interactions (refer to poster **Advanced Data Analysis Pipeline for Determination of Protein Complexes and Interaction Networks at the Genomics:GTL Center for Molecular and Cellular Systems**). MiPPI is linked to the CMCS laboratory information management system (LIMS) which maintains sample metadata, from cloning through MS analysis. The first public data release is scheduled for February 2007 and includes all center results collected through November 2006. This release contains biological and technical replicates of more than 300 bait proteins collected from over 500 pulldown experiments and over 900 mass spectrometric analyses. The database includes over 50,000 observed protein-protein interactions. Updates to MiPPI will be released semiannually.

Beginning in February 2007, the web interface (www.ornl.gov/sci/GenomestoLife/) to the MiPPI database will provide online searches by protein or protein-protein interaction, and will include a protein-protein interaction viewer for the observed interactions. Mass spectrometry results and corresponding metadata will be provided for download in mzXML and DTASelect file formats. Identified protein-protein interactions including the statistical analysis scores will be provided for download in delimited text file format.

43 ^{GTL}

Advances in Coverage and Quality for High-Throughput Protein-Protein Interaction Measurements

Jennifer Morrell-Falvey,¹ Mitchel J. Doktycz,¹ Dale A. Pelletier,¹ Linda J. Foote,¹ Elizabeth T. Owens,¹ Sankar Venkatraman,¹ W. Hayes McDonald^{1*} (mcdonaldwh@ornl.gov), Brian S. Hooker,² Chiann-Tso Lin,² Kristin D. Victry,² Deanna L. Auberry,² Eric A. Livesay,² Daniel J. Orton,² H. Steven Wiley,² and **Michelle V. Buchanan**¹

¹Oak Ridge National Laboratory, Oak Ridge, Tennessee and ²Pacific Northwest National Laboratory, Richland, Washington

Project Goals: The Center for Molecular and Cellular Systems (CMCS) focuses primarily on the objectives outlined in Goal 1 of the DOE Genomics:GTL program. The core of the CMCS is a high throughput pipeline for the identification of protein complexes, currently focusing on *Rho*-

rhodospirillum rubrum. The pipeline employs affinity isolation coupled with mass spectrometry to identify protein interactions. Computational tools are used to assess the significance of identified interactions. A dynamic research program supports the goals of the CMCS by focusing on the development and implementation of improved capabilities for complex isolation, molecular level identification of the complexes, and critical bioinformatics and computing capabilities. These efforts are focused on constructing a knowledge base that can provide insight into the relationship between the complement of protein complexes in these microbes and their biological function.

The overarching goal of the Center for Molecular and Cellular Systems (CMCS) is to identify protein interaction networks that form the molecular basis of biological function in microbes. To accomplish this goal, we have established a high-throughput analysis pipeline that is centered on generalized affinity-based isolation of protein interactors combined with mass spectrometric identification of the interacting protein components (refer to poster **Global survey of protein-protein interactions in *Rhodospirillum rubrum***). While this approach has proven successful for identifying a large number of protein interactions within the photoheterotrophic bacterium *Rhodospirillum rubrum*, interactions among some classes of proteins (e.g. membrane-associated or low abundance) remain difficult to detect. Approaches for overcoming these technological challenges as well as increasing throughput are needed for accomplishing our goal of identifying protein interaction networks with high confidence. Here we describe several strategies that we are developing to address these issues.

Our pipeline currently employs a tandem affinity tag comprised of 6XHis and the V5 epitope. Improvements in the affinity and specificity of the affinity tags used for labeling proteins will impact throughput by allowing the use of smaller culture volumes and improving detection of lower abundance proteins. In addition to improved affinity, the tags should be compatible with elution of the bound complex under nondenaturing conditions and amenable to automation using 96-well based robotic handling and microfluidic manipulations. For these reasons, we are constructing and testing several new Gateway-compatible vectors for expression of carboxy-terminal tags, including 1) 2X strep tag-6X His; 2) calmodulin binding protein (CBP)-3X FLAG; and 3) CBP-2X Protein A. In addition, two TEV protease sites are included in these constructs to facilitate more efficient elution from the first round of purification. To aid in detection and also potentially for use as another affinity capture moiety, these constructs also contain a tetracysteine tag that can be recognized by FIAHTM reagents. Calmodulin binding protein, which has reversible binding, and the Strep tag, which can be competitively eluted, were chosen based on the requirement for native elutions.

To facilitate reductions in sample amounts and to increase throughput, we are exploring optimizations to our high performance liquid chromatographic (HPLC) separations and our mass spectrometric data acquisition equipment and parameters. Comparisons are underway between a more traditional three dimensional ion trap (ThermoFinnigan LCQ) and a newer linear ion trap (ThermoFinnigan LTQ). These include comparisons between data acquisition rates, sensitivity, and effective dynamic range. HPLC optimizations are being performed in parallel in order to both optimize data acquisition duty cycles and take advantage of differences in speed between the two instruments. These parallel optimizations will provide increases in throughput while maintaining or even increasing the dynamic range and sensitivity of our analysis pipeline.

In addition to affinity isolation and mass spectrometric characterization, we can identify and confirm predicted protein interactions using a live cell imaging-based assay that exploits specific localization patterns in cells. This assay involves co-expression of two fusion proteins in *Escherichia coli*. The first protein of interest is directed to the cell poles by fusion to DivIVA and the second protein of interest is fused to GFP. A direct interaction between the two proteins results in recruitment of the GFP-fusion protein to the poles. Importantly, this assay can be used to test interactions among both soluble

and integral membrane proteins and is amenable to automation due to its rapidity, small scale, and ease of interpretation. To facilitate rapid analysis, we have also developed an automated image analysis algorithm to calculate the presence and location of GFP-fusion proteins in *E. coli* cells. Features such as cell number, diameter, area, and number of GFP-fusion protein localization sites are extracted from each image and used to relay quantitative values that aid in the scoring of positive interactions. This assay facilitates the directed analysis of protein interactions in live bacterial cells with the added benefit of amenability to automation.

Membrane-bound proteins are integral to many biologically active complexes, yet traditional isolation and purification of these proteins is usually tedious and inefficient. In order to make such purifications more routine, we have developed a co-fractionation strategy to localize and separate complexes under native states, followed by direct MS/MS analysis of the digested protein fractions. In this strategy we have tried to minimize dissociation between subunits, and thus loss of previously unknown subunits. For demonstration, clones of recombinant His-tagged ATP synthase were expressed in *Shewanella oneidensis* MR-1. Membrane proteins were solubilized and separated under native conditions in order of ionic strength on a Mono Q column. Fractions collected were trypsin digested and analyzed by LC MS/MS (Thermo-Finnigan LCQ). Results revealed that not only the ATP synthase subunits were eluted in common fractions, but also that proteins within other complexes were co-eluted at different ionic strengths, suggesting the presence of intact protein complexes. In parallel, we detected in-gel ATP hydrolysis approximately at the molecular size of the synthase complex (> 450 kD). Two dimensional electrophoresis images of the dissected gel show subunits ranging from 15 kD to 60 kD in size. These data demonstrate that co-fractionation and electrophoretic separation of membrane proteins coupled with mass spectrometric analysis is a valid and rapid way to analyze intracellular protein complexes.

Finally, in order to increase overall throughput, automation protocols for cloning, streaking, re-arraying, and purification steps are in place and automated affinity isolation protocols are being tested.

44 [—]GTL

Genome-Wide Identification of Localized Protein Complexes in *Caulobacter*

P. Viollier^{1*} (patrick.viollier@case.edu), J. Werner,² S. Pritchard,¹ E. Chen,² E. Huitema,¹ L. Shapiro,³ and Z. Gitai² (zgitai@princeton.edu)

¹Department of Molecular Biology and Microbiology, Case Western Reserve University, Cleveland Ohio; ²Department of Molecular Biology, Princeton University, Princeton, New Jersey; and

³Department of Developmental Biology, Stanford University, Stanford, California

Project Goals: With the availability of > 600 fully sequenced bacterial genomes, systematic approaches hold the key to exploiting our new found wealth of biological information. It is now clear that multienzyme complexes occupy discrete subcellular positions, suggesting also the existence of specialized catabolic or anabolic multi-enzyme complexes or enzymatic centers with other distinct metabolic functions. Towards the goal of assimilating a comprehensive 4D-topographic map of all co-localized proteins and protein complexes within a bacterial cell, we have embarked on a genome-wide fluorescence-based localization screen to determine the subcellular position of the finite set of proteins encoded in the genome of the Gram-negative, alpha-proteobacterium *Caulobacter crescentus*. In addition to determining the subcellular distribution of each

protein within the cell, the ability to synchronize *Caulobacter*, will allow U.S. to plot this 3D-data set into the fourth dimension (time), determining the dynamics of localization as cells progress through cell cycle. Two complementary strategies are underway by the Gitai (Princeton University) and Viollier (Case Western Reserve University) laboratories to achieve this objective.

With the availability of > 600 fully sequenced bacterial genomes, systematic approaches hold the key to exploiting our new found wealth of biological information. A major objective towards the feasibility of using bacterial cells in bioremediation and as alternate fuel sources is not only to outline protein-protein interactions on a genome-wide scale, but also determine where and when these interactions occur within the bacterial cell. It is now clear that multienzyme complexes occupy discrete subcellular positions, suggesting also the existence of specialized catabolic or anabolic multi-enzyme complexes or enzymatic centers with other distinct metabolic functions. These centers might provide the high enzymatic densities needed to facilitate maximum turn-over rates and/or they might also be important in channeling the substrates and products for ensuing reactions to the location where the reactions must take place. Towards the goal of assimilating a comprehensive 4D-topographic map of all co-localized proteins and protein complexes within a bacterial cell, we have embarked on a genome-wide fluorescence-based localization screen to determine the subcellular position of the finite set of proteins encoded in the genome of the Gram-negative, alpha-proteobacterium *Caulobacter crescentus*. In addition to determining the subcellular distribution of each protein within the cell, the ability to synchronize *Caulobacter*, will allow U.S. to plot this 3D-data set into the fourth dimension (time), determining the dynamics of localization as cells progress through cell cycle. Two complementary strategies are underway by the **Gitai** (Princeton University) and **Viollier** (Case Western Reserve University) laboratories to achieve this objective.

The Gitai laboratory is using the Gateway system from Invitrogen (www.invitrogen.com) to develop a library of fluorescently-tagged fusion proteins. There are several advantages to this approach: (1) All predicted orfs can be systematically labeled; (2) Since this is a targeted labeling scheme, we can prioritize the order of labeling and analysis of genes to investigate genes of particular interest first; (3) The number of proteins to screen for localization properties is constrained (about 3,800 for *Caulobacter*); (4) The Gateway system is modular—once entry clone are constructed, the gene of interest can be easily subcloned into a wide variety of Destination Vectors. To date, such entry clones have been generated for over 3200 genes (~85% of the predicted proteome). Roughly 2200 (~60%) of these genes have been mobilized into a destination vector that allows for their expression in *Caulobacter* as C-terminal mCherry fusions. The imaging and analysis of this first draft of the *Caulobacter* localizome is currently underway.

In a complementary approach, the Viollier laboratory has engineered random libraries of *Caulobacter* strains expressing protein or protein fragments fused to a fluorescent reporter (GFP or mCherry). These strains are being examined by fluorescence microscopy for fusions that are localized to specific cellular positions. There are several advantages to this approach: (1) It is unbiased; (2) It will produce localized protein fragments that will shed light on localization determinants in the complete protein; and (3) It is fast. After examining ~ 15,000 strains, approximately 0.2-3 % of the strains were found to express localized proteins, ranging from regulatory proteins to metabolic enzymes. These include several previously identified polarly-localized proteins, such as the CheR chemotaxis methylase, the DivL tyrosine kinase and the TipN polarity marker and metabolic enzymes (e.g amino acid and TCA cycle metabolism) that were also found to exhibit preferential localization to certain subcellular sites.

Follow-up studies are now underway in the Gitai and Viollier labs to elucidate how they are localized and the physiological consequences of mislocalizing them. Such studies have led to important advances on the physiological roles of these proteins and subcellular organization in *Caulobacter*

(Viollier et al. 2002; Viollier et al. 2002; Gitai et al. 2004; Dye et al. 2005; Gitai et al. 2005; Huitema et al. 2006; Kim et al. 2006). In addition, the localized fusion proteins generated by both of these approaches will subsequently be systematically analyzed using an automated mislocalization screen conducted by the **Shapiro** and **McAdams** laboratories (Stanford University) in an effort to isolate *trans*-determinants for localization. These studies complement our previous experiments that charted the chromosome layout and dynamics in live *Caulobacter* cells. Together our efforts present important strides towards defining the 4-dimensional macromolecular organization of a bacterial cell using schemes that are broadly applicable to any bacterial cell of interest.

References

1. Dye, N. A., Z. Pincus, J. A. Theriot, L. Shapiro and Z. Gitai (2005). "Two independent spiral structures control cell shape in *Caulobacter*." *Proc Natl Acad Sci U S A* 102(51): 18608-13.
2. Gitai, Z., N. Dye and L. Shapiro (2004). "An actin-like gene can determine cell polarity in bacteria." *Proc Natl Acad Sci U S A* 101(23): 8643-8.
3. Gitai, Z., N. A. Dye, A. Reisenauer, M. Wachi and L. Shapiro (2005). "MreB actin-mediated segregation of a specific region of a bacterial chromosome." *Cell* 120(3): 329-41.
4. Huitema, E., S. Pritchard, D. Matteson, S. K. Radhakrishnan and P. H. Viollier (2006). "Bacterial birth scar proteins mark future flagellum assembly site." *Cell* 124(5): 1025-37.
5. Kim, S. Y., Z. Gitai, A. Kinkhabwala, L. Shapiro and W. E. Moerner (2006). "Single molecules of the bacterial actin MreB undergo directed treadmilling motion in *Caulobacter crescentus*." *Proc Natl Acad Sci U S A* 103(29): 10929-34.
6. Viollier, P. H., N. Sternheim and L. Shapiro (2002). "A dynamically localized histidine kinase controls the asymmetric distribution of polar pili proteins." *EMBO J* 21(17): 4420-8.
7. Viollier, P. H., N. Sternheim and L. Shapiro (2002). "Identification of a localization factor for the polar positioning of bacterial structural and regulatory proteins." *Proc Natl Acad Sci U S A* 99(21): 13831-6.

45 ^{GT}

The Structure and Function of the *Caulobacter* MreB Actin-Like Cytoskeleton

N. Dye,^{1,2*} M. Mielke,³ Z. Pincus,² J. Theriot,² L. Shapiro,¹ and **Z. Gitai**³ (zgitai@princeton.edu)

¹Department of Developmental Biology, Stanford University, Stanford, California; ²Department of Biochemistry, Stanford University, Stanford, California; and ³Department of Molecular Biology, Princeton University, Princeton, New Jersey

Project Goals: The bacterial MreB cytoskeleton is an actin-like structure that represents an essential integrator of spatial and temporal cellular information. Despite its importance, however, the mechanisms by which MreB regulates processes such as cell morphogenesis, chromosome segregation, and protein localization remain unknown. Consequently, our objective is to determine the structure, function, and regulation of the MreB cytoskeleton. These studies will both illuminate broadly conserved processes that are essential to the survival of all bacterial species, and serve as a road-map for the mechanistic dissection of additional dynamic protein structures in bacteria.

The bacterial MreB cytoskeleton is an actin-like structure that represents an essential integrator of spatial and temporal cellular information. Despite its importance, however, the mechanisms by which MreB regulates processes such as cell morphogenesis, chromosome segregation, and protein localization remain unknown. Consequently, our objective is to determine the structure, function, and

regulation of the MreB cytoskeleton. These studies will both illuminate broadly conserved processes that are essential to the survival of all bacterial species, and serve as a road-map for the mechanistic dissection of additional dynamic protein structures in bacteria.

To probe the structure and organization of MreB, we are developing novel methods of analyzing its assembly and dynamics, both *in vivo* and *in vitro*. By imaging and tracking single molecules of fluorescently-labeled MreB, we have been able to visualize MreB kinetics in living cells (Kim et al. 2006). From these data we determined that MreB monomers treadmill through individual filaments in a polar manner, but that the polarity of each individual filament appears independent of the overall cellular polarity. We have also used these tracking data to model MreB's average filament length and polymerization and depolymerization rates Kim, 2006 #276}. These results suggest that the previously-observed MreB helical ultrastructures consist of multiple short, bundled filaments. To directly validate these models, we are now collaborating with the Ellisman (UCSD) and Larabell (LBNL) groups to develop EM- and XM-compatible strategies for ultra-high-resolution-imaging of MreB and MreC structures.

To identify proteins that function with MreB, we first took a candidate approach, examining proteins that are co-conserved with MreB. By combining mutant and localization analysis, we determined that the cytoplasmic MreB cytoskeleton functions in parallel to a periplasmic helical structure made up of the MreC protein (Dye et al. 2005). By developing a novel computational method for analyzing cell shapes, we also determined that MreB and MreC collaborate to determine proper morphology by regulating the proper localization of a transmembrane peptidoglycan rearrangement enzyme, Pbp2 (Dye, Pincus et al. 2005). Our shape analysis software should be widely applicable to other researchers studying irregular cell morphologies.

Currently, we are using several genetic and biochemical approaches to identify and characterize additional upstream regulators and downstream mediators of MreB function. We have succeeded in assembling filaments of purified MreB *in vitro*, and we are now characterizing the mechanistic details of MreB polymerization. Such *in vitro* polymerization of fluorescently tagged MreB is providing a method to assay candidate MreB regulators. Together, these studies will help U.S. understand a fundamental and conserved macromolecular complex, while serving as a platform for the development of new methods that will be broadly applicable to other bacterial systems of interest to the DOE.

References

1. Dye, N. A., Z. Pincus, J. A. Theriot, L. Shapiro and Z. Gitai (2005). "Two independent spiral structures control cell shape in *Caulobacter*." *Proc Natl Acad Sci U S A* **102**(51): 18608-13.
2. Kim, S. Y., Z. Gitai, A. Kinkhabwala, L. Shapiro and W. E. Moerner (2006). "Single molecules of the bacterial actin MreB undergo directed treadmilling motion in *Caulobacter crescentus*." *Proc Natl Acad Sci U S A* **103**(29): 10929-34.

EM Tomography Enhancements

Fernando Amat^{1*} (famat@stanford.edu), Farshid Moussavi^{1*} (farshid1@stanford.edu), Kenneth H. Downing,³ Mark Ellisman,⁴ Luis R. Comolli,³ Albert Lawrence,⁴ Mark Horowitz,^{1*} and **Harley McAdams**² (hmcadams@stanford.edu)

¹Electrical Engineering Department and ²Developmental Biology Department, Stanford University, Stanford, California; ³Lawrence Berkeley National Laboratory, Berkeley, California; and ⁴National Center for Microscopy and Imaging Research, University of California, San Diego, California

Project Goals: Development of high-throughput methods to identify and characterize spatially localized multiprotein complexes in bacterial cells.

Electron microscope tomography is a powerful tool for studying the 3D structure and organization of biological specimens. However, it presents particular challenges for attaining high enough resolution. One such limitation is the inability to collect high resolution images of sections at high tilt due to apparent specimen thickening that occurs as they are tilted. This effect produces blurred images due to chromatic aberrations at high tilt, in addition to the so-called “missing wedge” of data, which together contributes to a significant reduction in resolution of tomographic reconstructions, especially in the z-axis. We have been pursuing several approaches in order to reduce these effects, including development of most-probable loss (MPL) energy filtering electron tomography [1], but are still ultimately constrained by the specimen geometry. To augment this MPL strategy we have developed a method allowing the fabrication of “prismatic” sections of resin embedded microbial specimens for ultra high tilt electron tomography. The procedure for preparing these and the associated tomographic volumes will be described.

Another obstacle to high resolution is alignment accuracy. This is especially true in the case of cryo EM tomography, which enables the study of cells in close to their “native” environment. Each image in a cryo EM tilt series is quite noisy, since the total electron dose through the tomographic tilt series must be constrained to limit structural damage to the cell. Even with gold markers added to the sample, robust automatic alignment of the cryo EM image data for reconstruction remains difficult, and manual intervention is required. We address this problem by leveraging recent work in probabilistic analysis, and have constructed a prototype alignment system using *Markov random fields* (MRF's) and robust optimization for alignment of tilt-series. This fully automatic alignment will become more critical as faster and better systems for automatic data acquisition are being developed. Our goal is to eliminate the alignment bottleneck in tomographic imaging. We are also integrating these methods with recent approaches that allow for correction of projection errors due to specimen changes during a tilt series and distortions due to properties of the electron optics [2].

With markers, there are three basic steps required to align the cryo EM dataset: marker feature identification, correspondence and tracking of these features throughout the image set, and projective model estimation from these feature tracks. Typically, automatic tracking makes many errors, which in turn cause the projection model estimation to fail. In our framework, we focus on reduction of tracking errors through the use of contextual information, as well as making the projective model estimation robust to any remaining tracking errors. Feature correspondence and tracking are accomplished at local and global levels. Local correspondence is between two images, and is accomplished by treating the set of markers as a Markov random field (MRF), that is, a set of variables that depend on each other only through their neighbors. We use mutual information and the relative geometric positions of pairs of markers to set the initial estimates for local correspondence, and use a standard

loopy belief propagation algorithm to estimate the most likely correspondences between image pairs. Global correspondence is achieved by combining the results of local correspondence in a tree-based comparison scheme with redundancy to form robust track estimates.

Errors in the tracks are possible due to feature location mistakes as well as inaccurate inference results. Therefore, the projective model estimation uses a robust fitting method as opposed to least squares that is tolerant to outliers. After we have an estimate of the projective model, the model is iterated using expectation maximization (EM) to re-estimate perceived outliers with improved reprojection data from the current model. This iteration is performed as many times as necessary before a stopping criterion is satisfied. In sample cases, we find that only a small number of iterations is needed (often only one).

This robust alignment framework has allowed U.S. to fully automatically recover dozens of contours (both complete and piecewise) with subpixel accuracy from several challenging cryo datasets of the bacterium *Caulobacter crescentus*. The results were used to create 3D reconstructions comparable to results previously obtainable only by extensive manual intervention.

All the techniques described above are implemented in the software package Robust Alignment and Projection Estimation for Tomographic Reconstruction (RAPTOR). This software will be made generally available.

References

1. Bouwer et al. (2004) J Struct Biol. 148(3):297-306.
2. Lawrence et al. (2006) J Struct Biol. 154(2):144-67.

47 ^{GT}

Automated Screen for Identification of Mislocalization and Morphological Mutants

G. Bowman, N. Hillson, M. Fero, S. Hong, L. Shapiro, and **H. H. McAdams*** (hmcadams@stanford.edu)

Department of Developmental Biology, Stanford University, Stanford, California

Project Goals: Development of high-throughput methods to identify and characterize spatially localized multiprotein complexes in bacterial cells.

Many dynamically localized bacterial proteins have been identified over the last few years by fluorescence microscopy, and in many of these cases the localization is essential for proper cell function. The number of localized proteins will vary across organisms. In *Caulobacter crescentus*, roughly 10% of expressed proteins are localized at some time in the cell cycle. Although the functions of many localized proteins have been extensively studied, there is limited understanding of the pathways and dependencies that produce localization or of the protein domains required for localization. Identification of these pathways and mechanisms involves laborious characterization of mutant strains. This is traditionally a “low throughput” process, particularly where the mutant phenotypes can only be detected by microscopic examination. We have developed a robotic screen to massively accelerate identification of mislocation mutants using automated microscopy and computational image analysis.

The localization mutants identified are then characterized and analyzed by conventional genetic and biochemical methods.

Our automated, high-throughput screen process can identify mutant strains that mislocalize both essential and non-essential proteins that normally have a distinctive spatial or temporal localization pattern. Development of this screening technology is a collaboration between the McAdams and Shapiro labs at Stanford, the Viollier lab at Case Western, and the Gitai laboratory at Princeton.

There are two stages to the screen process. First, we identify the proteins that exhibit a temporal or spatial localization pattern and characterize that localization pattern. This is being done in collaboration with the Viollier laboratory at Case Western University and the Gitai laboratory at Princeton University. Second, for a subset of the localized proteins, we use genetic screening to identify mutations that cause localization defects (in timing or placement). For each target localized protein, we generate strains with location defects. We use transposon mutagenesis for the nonessential genes. The most challenging mutants, of course, are those targeting proteins whose proper location is essential. We generate temperature-sensitive mutants for these strains. Location defects may involve complete failure to localize or localization at the wrong time or place. The screen will identify both intragenic mutations that indicate portions of the protein required for its cellular localization, and extragenic mutations in genes encoding interacting proteins required for either specific or general localization processes.

We use the automated microscopy system to examine thousands of the mutagenized strains to identify those with location defects. We first spot the mutagenized strains on a novel microscope slide design to obtain an array of samples on small agar pads under a cover slip. Fluorescent micrographs of the resulting cell sample arrays are taken rapidly with the computer controlled microscope. The rectangular arrays of agar pads provide a specific row-column address on the slide that is correlated with the source wells in the 96-well storage. We have developed automatic image processing algorithms to find, orient, measure, and classify cells in the microscopic images. This brute force, yet totally automated, system for finding localization mutants is generally applicable to any protein whose GFP-tagged variant shows a distinctive localization pattern in the cell.

An integral part of our screen methodology involves detailed computational analysis of the size and shape of collections of cells in the microscope image field. The program we developed for the image analysis automatically extracts individual cell images from a field of cell images. The properties of the cell shapes and localized protein positions are then used to classify the cells according to stage of cell development and localization pattern. The characteristics of the sample cells in the mutagenized sample are then compared to characteristics of wild type cell populations. Sorting by cell cycle stage is necessary because many proteins are transiently localized at a distinct time in the cell cycle. Although our primary focus is on identification of mislocation mutants, the image classification analysis also identifies mutant strains with defects in cell shape or in cell cycle progression. This enables identification of cytoskeleton or cell decision defects as a valuable byproduct result.

The key requirement for the screen is the ability to examine very large numbers of mutagenized strains by fluorescent microscopy. Key characteristics of the fluorescent microscope and associated acquisition software are automated X-Y-Z stage control, automated switching from transmission microscopy modes such as phase contrast or DIC to fluorescence imagery, automated time lapse imaging, and fully automatic focusing. In addition, the temperature of the sample and stage are monitored and controlled via an environmental enclosure. Images are automatically acquired using a megapixel cooled CCD camera and transferred to an image repository where they are queued for image analysis. Automation of computer control enables thousands of strains to be screened rapidly with minimal manual involvement in the microscopy. We can use a loose tolerance for false positives

in the automated image analysis, since the cell images that the computer analysis categorizes as probable mutants are then examined by eye.

For our initial mutant search, we constructed a strain with three localized proteins labeled with distinct fluorescent tags: PleC, which localizes to the swarmer pole, was labeled with YFP; DivJ, which localizes to the stalked pole, was labeled with mCherry; and ZapA, which localizes to the division plane, was labeled with CFP. The use of the multi-labeled strain enabled a simultaneous search for strains with mislocation of any of the three labeled proteins, so that the mutant yield is higher.

48 [—]_{GTL}

Methods for *in vitro* and *in vivo* Imaging of Protein Complexes

Huilin Li* (hli@bnl.gov), James Hainfeld* (hainfeld@bnl.gov), Minghui Hu, Michael Mylenski, Kevin Ryan, Luping Qian, Raymond P. Briñas, Elena S. Lyman, and Larissa Kusnetsova

Biology Department, Brookhaven National Laboratory, Upton, New York

Project Goals: To develop new high resolution imaging method and to apply the method for imaging protein complexes in biological systems that is relevant to DOE mission.

The structures of biological molecular assemblies and their locations inside the cell are keys to understanding the function. Our *in situ* bi-mode cryo-electron tomography and site-specific labeling method, which takes advantage of ultra-structural visualization capability of the cryo-TEM and the heavy metal cluster label detection capability of the cryo-STEM, holds the potential for simultaneous three-dimensional structural visualization and protein mapping.

Developing the cryo-STEM tomography capability on a commercial Jeol 2010F cryo-TEM/STEM microscope.

Our approach requires low dose imaging and tomographic capability in both TEM and STEM modes. A modern commercial transmission electron microscope comes with these capabilities only in TEM mode, but not in STEM mode. In last year, we developed a Gatan Digital Micrograph Plug-In that we called it STEMan that enable low dose image acquisition. During this year, we added STEM tomography capability, and the program is now called AuotSTEM for automated scanning transmission electron microscopy tomography (Fig. 1). AutoSTEM was based on scripts created using Digital Micrograph (DM), since we use the Gatan STEM acquisition device DigiScan. Jeol's FasTEM Communication Kit (FTCOMM) and Digital Micrograph's Software Developers Kit (DMSDK) were essential in creating new functions to implement auto-tracking and auto-focusing. Overall, we developed five DM scripts and one Microsoft's Visual Studio-based dynamic link library (DLL). The DLL communicates between the microscope and DM. The Auto-focusing part utilized an image-gradient based algorithm. The computation for auto-tracking and auto-focusing is primarily based on the DM scripts. The program is completed but further testing is needed.

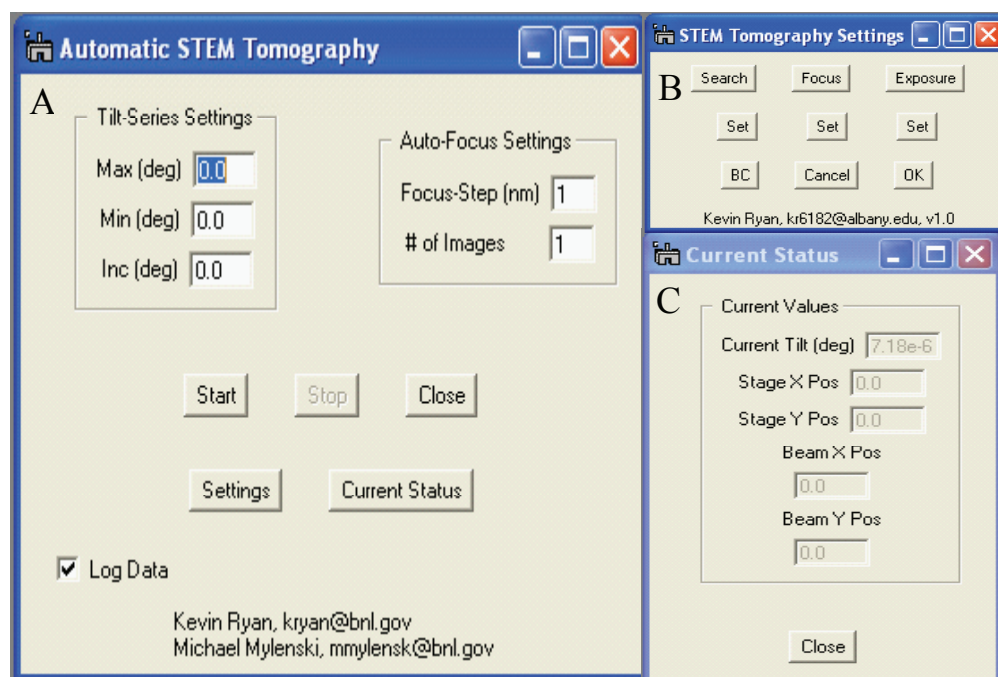


Fig. 1. GUIs for AuotSTEM. (A). The main Graphical User Interface (GUI) in Digital Micrograph for automated STEM tomography. It controls the auto-tilting, auto-focusing parameters, the starting and stopping of the tilt series, and access to the settings and current status GUIs. (B). GUI for adjusting microscope settings for auto-tracking and auto-focusing. (C) GUI when invoked displays the current microscope settings, such as specimen tilt angle, the current stage and current beam positions.

Nanoparticle assembly of bacterial proteins and complexes and electron microscopy – Potential use for alternative energy production.

We have used site-specific binding gold nanoparticles to assemble proteins and complexes into higher order structures that could be useful for alternative energy production and visualized them by EM. For example, redox enzymes can oxidize ethanol to produce free electrons which, if harnessed, would directly produce electricity. In order to do this, the enzymes must be assembled on an electrode and the flow of electrons efficiently coupled to it. This was shown to be possible for glucose oxidase using gold nanoparticles: the electron would normally be received by O_2 in solution, but could be routed to a metal electrode by attaching the enzyme and providing a conductive path via the nanoparticle to the metal surface (Xiao Y, Patolsky F, Katz E, Hainfeld JF, Willner I: Plugging into Enzymes: nanowiring of redox enzymes by a gold nanoparticle. *Science*, 299, 1877, 2003.).

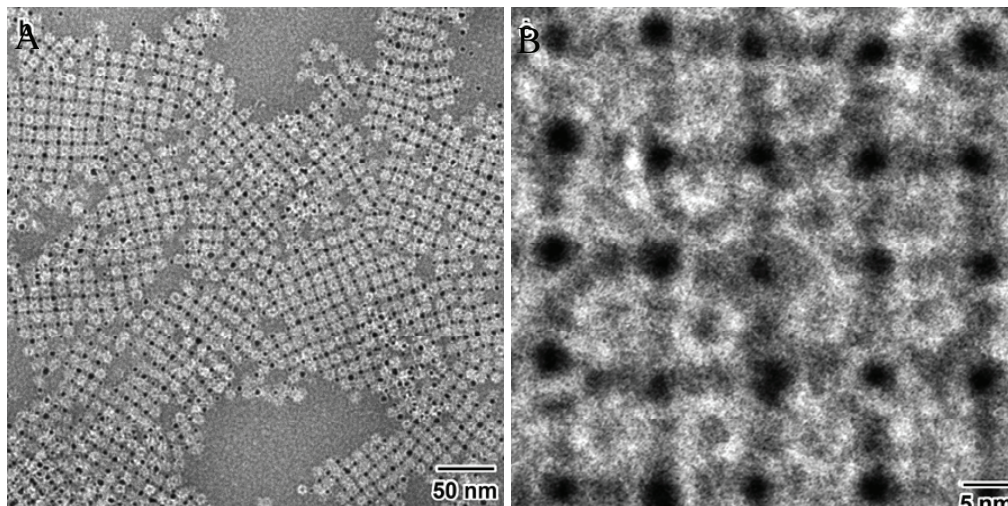


Fig. 2. The bacterial 20S proteasomes organized into 2-D arrays using 4.4 nm gold nanoparticles (black dots). (A) low mag and (B) high mag.

One obstacle to producing efficient enzyme electrodes is to coat the electrode with the redox enzyme in a functional monolayer. Here we have taken bacterial enzyme complexes and shown by electron microscopy that large monolayers could be assembled by use of gold nanoparticles functionalized to “glue” the enzymes together into an oriented and ordered sheet. This sheet was shown to be strong enough to transfer onto a carbon film, the EM grid (Fig.2). The protein used was *Mycobacterium tuberculosis* 20S proteasomes which were expressed with a 6x-His tag. They were organized into arrays by binding to multifunctional 4.4 nm gold nanoparticles derivatized with the nickel nitrotriacetic acid group (Ni-NTA). **Using this newly developed nanotechnology, it should be possible to create ordered redox enzyme electrodes leading to more efficient energy conversion to electricity of biofuels.**

Acknowledgement: This research is supported by a grant from the Office of Biological and Environmental Research of the U.S. Department of Energy (KP1102010).

49 ^{GT}L

Probing Single Microbial Proteins and Multi-Protein Complexes with Bioconjugated Quantum Dots

Gang Bao^{1*} (gang.bao@bme.gatech.edu), Grant Jensen,² Shuming Nie,¹ and Phil LeDuc³

¹Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, Georgia; ²Department of Biology, California Institute of Technology, Pasadena, California; and ³Department of Mechanical Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania

Project Goals: The goal of this DOE/GTL project is to develop quantum-dot (QD) based strategies for imaging and identification of individual proteins and protein complexes in microbial cells with high specificity and sensitivity. The multifunctional quantum-dot bioconjugates consist of quantum dots of 2-6 nm in size encapsulated in a phospholipid micelle, and delivery peptides and protein targeting ligands conjugated to the surface of the QDs. Once the QD bioconjugates are

internalized into microbial cells by the peptide, the adaptor molecules on the QD surface bind to specific target proteins or protein complexes that are genetically tagged. Optical imaging is used to visualize the localization, trafficking and interaction of the proteins, resulting in a dynamic picture but with a limited spatial resolution (~ 200 nm). The same cells are then imaged by EM to determine their detailed structures and localize the target proteins to ~ 4 nm resolution. For each protein or protein complex, selected tags will be tested to optimize the specificity and signal-to-noise ratios. This innovative molecular imaging approach integrates peptide-based cellular delivery, protein targeting/tagging, light microscopy and electron microscopy, and will have a wide range of biological applications relevant to the GTL program at DOE.

We have been developing quantum-dot (QD) based strategies for imaging and identification of individual proteins and protein complexes in microbial cells. Currently, there is a lack of novel labeling reagents for visualizing and tracking the assembly and disassembly of multi-protein molecular machines. There is no existing method to study simultaneous co-localization and dynamics of different intra-cellular processes with high spatial resolution. As shown in Figure 1, the multifunctional quantum-dot bioconjugates we develop consisting of a quantum dot of 2-6 nm in size encapsulated in a phospholipid micelle, with delivery peptides and protein targeting ligands (adaptors) conjugated to the surface of the QD through a biocompatible polymer. After internalization into microbial cells, the adaptor molecules on the surface of QD bioconjugates bind to specific target proteins or protein complexes that are genetically tagged. Optical imaging is used to visualize the localization, trafficking and interaction of the proteins, resulting in a dynamic picture but with a limited spatial resolution (~ 200 nm). The same cell is imaged by EM to determine their detailed structures and localize the target proteins to ~ 4 nm resolution. For each protein or protein complex, selected tags are tested to optimize the specificity and signal-to-noise ratios of protein detection and localization. This innovative molecular imaging approach integrates peptide-based cellular delivery, protein targeting/tagging, light microscopy and electron microscopy.

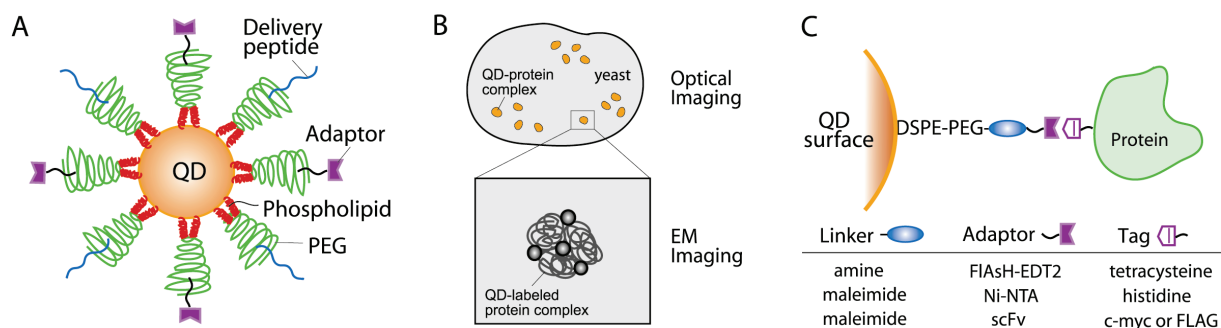


Figure 1. (A) Schematic illustration of a multifunctional quantum dot bioconjugate consisting of encapsulated QD with targeting adaptor and delivery peptide on its surface; (B) correlated optical and EM imaging of the same cell gives both temporal and spatial information on a protein complex; (C) possible conjugation and tagging strategies for optimizing detection specificity and sensitivity. Note that molecules are not drawn to the exact scale.

To achieve the goals of this DOE GTL project, we have synthesized core-shell and alloyed quantum dots (QDs) for dual-modality optical and EM imaging. This new class of QDs contains Hg, a heavy element that is often used in x-ray and electron scattering experiments, allowing studies of cellular structures at nanometer resolution. We have also linked QDs to a chelating compound (nickel-nitrilotriacetic acid or Ni-NTA) that quantitatively binds to hexahistidine-tagged biomolecules with controlled molar ratio and molecular orientation.

To target semi-conductor quantum dots (QDs) to specific intracellular proteins, we constructed fusion proteins including the commercially available SNAP tag. The SNAP tag is a 22 kDa protein that is an engineered form of the human O6-alkylguanine-DNA-alkyltransferase (AGT). The SNAP protein is also able to transfer residues from para-substituted benzylguanines, resulting in the covalent attachment of the substituted group (such as a QD) to the SNAP tag. As a model system, SNAP-DsRed-Monomer-Actin fusion proteins were expressed in *E.coli* and purified in order to perform *in vitro* polymerization experiments to prove the functionality of the fusion proteins, which was assessed both by polymerizing actin and then adding QDs for tagging of the filaments and by attempting to polymerize the actin already labeled with the QDs. The use of the DsRed-Actin fusion allowed for continuous monitoring of both the actin and the QDs without the need for fixing and staining. In addition, the pore-forming bacteriotoxin streptolysin-O (SLO) was used to deliver benzylguanine-conjugated QDs into 3T3 cells transfected with pSNAP-DsRed-Monomer-Actin. The feasibility of using this covalent method to label intracellular proteins in live cells was assessed.

As part of our effort to develop QD-based technologies to identify and track individual protein complexes in microbial cells, we have performed preliminary optical imaging studies of single QDs delivered into living cells. Using a spinning-disk confocal microscope, we have succeeded in imaging single QD probes delivered into the cytoplasm of living cells. Several lines of evidence support that the QDs in cells are indeed single: (a) these QDs have similar brightness and spot size; (b) the brightness of these QDs is not higher than that of single QDs on a coverslip; and (c) the intracellular QDs show intermittent on/off light emission (called blinking), a characteristic of single dot behavior. We have also developed computation algorithms for two-color colocalization and correlation tracking of QD probes. As an alternative, we successfully imaged individual 10 nm gold nanoparticles and established the darkfield optical imaging capability for cellular studies.

We are advancing electron tomography as a promising new tool to image protein complexes both *in vitro* and *in vivo* within small microbial cells. A new helium-cooled, 300kV, FEG, “G2 Polara” FEI TEM at Caltech was used to image purified protein complexes, viruses, and whole bacterial cells. We pioneered the use of a new “flip-flop” cryorotation stage that allows dual-axis cryotomography, and developed a simple Perl-based system for distributed computation to handle the massive image processing demands that arise from imaging intact bacteria in 3D. These technological advances have allowed U.S. to visualize directly cytoskeletal elements within small microbial cells and the domain structure of purified multienzyme complexes, both are key imaging goals of the Genomics:GTL program. For example, using electron cryotomography of whole cells, we revealed the *in situ* structure of the complete flagellar motor from the spirochaete *Treponema primitia* at 7 nm resolution. Twenty individual motor particles were computationally extracted from the reconstructions, aligned and then averaged. The stator assembly, revealed for the first time, possessed 16-fold symmetry and was connected directly to the rotor, C ring and a novel P-ring-like structure. The unusually large size of the motor suggested mechanisms for increasing torque and supported models wherein critical interactions occur atop the C ring, where our data suggest that both the carboxy-terminal and middle domains of FliG are found.

We have successfully integrated quantum dots into *Dictyostelium* through culturing them simultaneously with bacteria, on which they are feeding. After incubation and both fluorescent and confocal microscopy imaging, the *Dictyostelium* reveal a distribution of quantum dots indicating that they are dispersed throughout the cell cytoplasm. This is compared with the control cells that have not been incubated with quantum dots and further with cell studies where endosomes reveal aggregated fluorescent patterns at earlier time periods. We have shown specific live cell labeling of *Dictyostelium* tagged with quantum dots that are targeted for F-actin using phalloidin. The fluorescent patterns are similar to the patterns for *Dictyostelium* that is immunofluorescently labeled with FITC-phalloidin.

Acknowledgement: This research is funded by a grant from DOE (DE-FG02-04ER63785).

50 ^{GTL}**Correlated Light and Electron Microscopy of Protein Complexes in *Caulobacter crescentus***

Guido M. Gaietta^{1*} (ggaietta@ncmir.ucsd.edu), Thomas J. Deerinck,¹ Grant Bowman,² Yi Chun Yeh,² Luis R. Comolli,³ Lucy Shapiro,² Harley McAdams,² and **Mark H. Ellisman**¹

¹National Center for Microscopy and Imaging Research, ^{*}Department of Neurosciences, University of California, San Diego, California; ²Department of Developmental Biology, Stanford University, School of Medicine, Beckman Center, Stanford, California; and ³Life Science Division, Lawrence Berkeley National Laboratory, University of California, Berkeley, California

Project Goals: Adaptation of a molecular tagging approach to correlated light and electron microscopy for 3D analysis of protein complexes in *Caulobacter crescentus*.

Many regulatory proteins and protein/DNA complexes in *Caulobacter crescentus* are found in specific locations with noted variation related to the stage in the cell cycle. A main objective of the *Dynamic spatial organization of multi-protein complexes controlling microbial polar organization, chromosome replication, and cytokinesis* GTL project is to adapt methods for correlated LM and EM imaging to the analysis of these structures in *Caulobacter*. Our efforts have been focusing on the use of the tetracycline/biarsenical molecular tagging system, developed in collaboration with Professor Roger Tsien's group at UCSD. Recombinant probes for proteins in key subsystems of *Caulobacter* are being generated and used as test bed to develop this approach, which allows U.S. to achieve optimal preservation of the ultrastructure, sensitive localization of the target complexes and their visualization in 3D reconstructions. The following table presents some of the strains currently under analysis:

Protein	Tag	Localization
FtsZ	GFP-4C, YFP-4C	Division plane
FtsK	GFP-4C	Division Plane
HU	GFP-4C	Chromatin
MipZ	GFP-4C, YFP-4C	Cell pole
MreC	4C	Cytoskeleton
MreB	4C	Cytoskeleton
LacI	GFP-4C, CFP-4C	Chromatin
McpA	GFP-4C	Cell Pole
CpaF	4C	Cell Pole
GmpA (Snc04)	GFP-4C	Cell Pole

As displayed in the table, recombinant forms of the target proteins often carry combinations of intrinsically fluorescent proteins (FP; GFP and its derivatives) tagged at the carboxyl terminus with the small tetracycline motif FLNCCPGCCMEP (4C). The resulting protein is visible at the light level by direct excitation of FP and, upon labeling of the tetracycline motif, of the biarsenical compounds FAsH and ReAsH. ReAsH is also used to trigger the photoconversion of diaminobenzidine (DAB) by either direct excitation or by FRET from the neighboring FP, hence generating a precipitate visible at the electron microscope (upon treatment with osmium tetroxide).

We are currently developing the labeling protocol using the pXyl-GmpA strain, in which we can achieve robust expression of tetracycline tagged GmpA protein using a multi-copy cassette under control of the xylose-dependant promoter pXyl. GmpA is a newly identified protein that localizes to the cell poles throughout the cell cycle of *Caulobacter*, and is essential for normal cell division.

When we induce high levels of protein expression in the pXyl-GmpA strain, the GmpA protein forms a large plug at the cell pole that displaces the normal cytoplasm. This enlarged region of GmpA accumulation has the effect of increasing the area of the cell pole, as the localization patterns of other polar proteins are extended to co-localize with the GmpA plug.

Using these enlarged polar features, we have shown that it is possible to apply the ReAsH labeling method to observe protein localization at high resolution at the cell poles. To do this, the labeling process was adapted from that used in eukaryotic systems, and streamlined so that *Caulobacter* cells could be easily stimulated, stained and prepped for photoconversion over a ~4 hour-long period. After stringent washes with a competing dithiol agent, ReAsH-labeled bacteria were pelleted, re-suspended and plated onto polyethylenimine-coated Mat-Tek dishes and, for optimal preservation of the bacterium ultrastructure, fixed in a glutaraldehyde/acrolein mixture. The photoconversion reaction is driven by exciting ReAsH by FRET from GFP, hence increasing the overall absolute contrast of the biarsenical, bringing it closer to that of GFP. Furthermore, FRET excitation excluded non-specific biarsenicals because of its tight space limitation between donor and acceptor (they must be no more than 0.8 nm apart), improved the actual specificity of photoconversion and provided a highly concentrated and reproducible deposition of DAB polymers at the cell pole, where GmpA locates. Very recent advancements in our technique have increased labeling efficiency, and we are making progress towards the goal of observing GmpA localization at normal expression levels. Eventually, we will compare the localization patterns of several polar proteins to determine the three dimensional structural organization of this important cellular domain. Applying this technique to proteins that have other kinds of localization patterns (table 1) will eventually allow U.S. to observe chromosome organization, division plane assembly, cytoskeletal proteins, and more.

In order to obtain an accurate, high-resolution map of the organization of proteins in *Caulobacter*, it is imperative to optimize the preservation of cellular ultrastructure. We find that we can greatly enhance preservation by combining the aldehyde/acrolein cross-linking properties and High Pressure Freezing (HPF) and freeze substitution. This hybrid method allows the introduction of a photoconversion step to deposit DAB for subsequent osmification during the low-temperature freeze-substitution process. Electron Tomography proceeds with the resin embedded samples with preservation comparable to that obtained by HPF from live specimens in absence of chemical cross-linking.

51 ^{GTL}

Computational Analysis of the Protein Interaction Networks of Three Archaeal Microbes

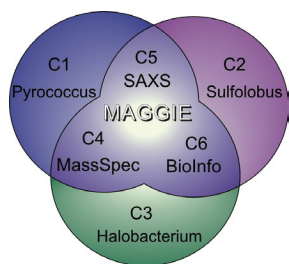
Chris Ding* (chqding@lbl.gov), Chunlin Wang, and **Stephen R. Holbrook**

Lawrence Berkeley National Laboratory, Berkeley, California

Project Goals: MAGGIE will address immediate GTL goals by accomplishing the following three overall goals: 1) provide a comprehensive, hierarchical map of prototypical microbial protein interactions, 2) develop and apply SAXS technologies for high throughput characterizations of protein conformations, shapes and assemblies controlling microbial cell biology, and 3) create and test powerful computational descriptions for the interpretation and eventual control of protein interactions and activities controlling microbial processes. We propose to develop

and implement a framework for integration, representation and analysis of proteomic data from both experimental and bioinformatics sources. This multi-scale representation proceeds from protein motifs to domains, proteins, complexes, super-complexes and pathways.

The computational identification and mapping of protein complexes and networks using genomic and experimental data is critical to our understanding of complex biological systems. In relation to the MAGGIE project to characterize the molecular machines and interactions of archaeal genomes, we have developed a software suite for analysis of their protein interaction network. These algorithms identify missing links in the network by transitive closure, extracted protein complexes as cliques, placed them into context by a hierarchical decomposition process using a distance metric and allowed for proteins to be shared between complexes. Finally, we have developed an algorithm for extracting bicliques from a data matrix and applied it to finding bicliques between protein domains and protein complexes.



Because of the abundance of experimental data, we have used yeast in development and testing of these programs, but as experiments proceed on microbial proteomes, we have begun to apply these approaches to *Pyrococcus furiosus*, *Sulfolobus solfataricus*, and *Halobacterium* NRC-1. An advantage to studying these three diverse archaea is the capability to characterize conserved components and organism specific systems of the microbial domain. This software and results drawn from analysis of these organisms will be made available to the GTL community.

52 ^{GTL}

The Use of Small Angle X-ray Scattering to Extract Low Resolution Structures and Monitor Sample Quality from Archeal Proteomes

Greg Hura^{1*} (glhura@lbl.gov), Michal Hammel,¹ Susan Tsutakawa,¹ Cesar Luna-Chavez,¹ Robert Rambo,¹ Ferris Poole,³ Francis Jenney,³ Angeli Lal Menon,³ Mike Adams,³ and **John Tainer**^{1,2}

¹Lawrence Berkeley National Laboratory, Berkeley, California; ²Skaggs Institute for Chemical Biology, The Scripps Research Institute, La Jolla, California; and ³Department of Biochemistry & Molecular Biology, University of Georgia, Athens, Georgia

Project Goals: Characterization of Archeal Proteomes

Small Angle X-ray Scattering (SAXS) is a high through-put technique for determining low resolution structures of macromolecular complexes. Ideal samples for SAXS data collection are 15-20 µg/µL of monodispersed and homogeneous macromolecules in solution. Samples which do not fit these criterion are identifiable from the SAXS signal and therefore SAXS may be used as a quality check on various preparations of a particular macromolecule. Samples which do fit this criterion produce scattering curves which may be used to define dimensions above 10 Å.

The SIBYLS beamline at the Advanced Light Source in Berkeley has been developed to conduct high through-put SAXS. Using a pipetting robot 800 samples may be collected in 40 hours. We intend to characterize native and recombinantly purified samples from organisms of GTL interest. By screening various buffer conditions we will determine stabilizing conditions for particular complexes. By screening libraries of compounds we will determine which metabolites cause con-

formational changes in various macromolecules. We will also develop and maintain a web based data base where low resolution structures of complexes are accessible to the scientific community.

A first pass has been conducted on 17 *Pyrococcus Furiosus* (Pf) recombinant proteins which were previously prepared as part of a crystallographic structural genomics initiatives. Three such samples were successfully crystallized in that effort while 14 remained structurally uncharacterized. Of the 17 samples collected 13 were immediately amenable for shape reconstruction, the other 4 were aggregating in solution. Various preparative conditions were attempted with one of the 4 aggregates which identified a successful non-aggregating preparation.

The low resolution shapes for the 3 samples which had a crystallographically determined structure showed excellent agreement with their structures. Within the remaining 10, for which low resolution structures were determined, large multimers were identified. Many of the structures have large flexible tails which likely complicate crystallographic studies. Proteolysis fragments may be more amenable. Some of the complexes have no homology to proteins with known function. Further biochemical characterization of the most interesting shapes will continue.

53 GTL

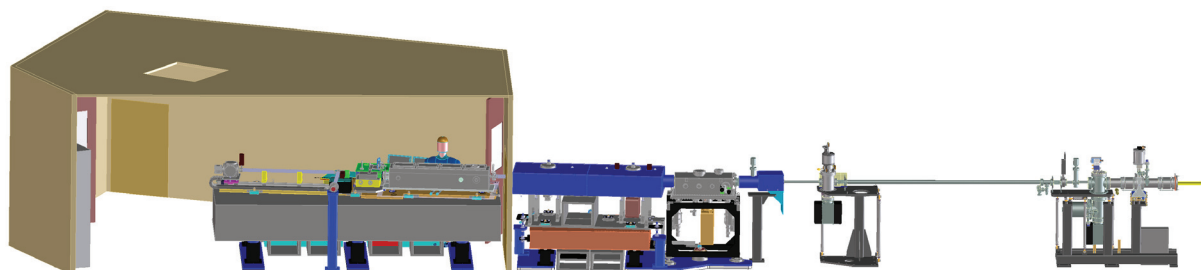
Imaging *Caulobacter crescentus* using Soft X-Ray Tomography: A New Imaging Tool for Genomics:GTL and Bioenergy Research

B.M. Maguire,^{1,3} C.A. Tonnessen,^{1,3} G. McDermott,^{1,3} A.J. McDonnell,^{1,3} M.A. Le Gros,^{2,3} and **C.A. Larabell^{1,2,3*}** (CALarabell@lbl.gov)

¹Department of Anatomy, University of California, San Francisco, California and ²Physical Biosciences Division and ³National Center for X-ray Tomography, Lawrence Berkeley National Laboratory, Berkeley, California

Project Goals: Charting cellular sub-structures and localizing proteins and multi-protein complexes in whole hydrated cells using X-ray tomography.

Developing a predictive systems-level understanding of an organism requires integration of data from a number of large scale ‘-omics’ programs, such as proteomics, metabolomics, and structural genomics. An unambiguous interpretation of this data also requires detailed knowledge of the cellular locations where the molecular interactions occur during the cell cycle, or when the organism is responding to specific environmental conditions. Soft x-ray tomography is a new tool for imaging cells and for localizing labeled proteins. This new imaging technique has an inherent advantage over established imaging techniques, such as electron microscopy, in that it only requires the use of simple sample preparation protocols and results in images with a spatial resolution significantly higher than can be obtained with light microscopy. The technique also has a further advantage in that soft x-rays images reveal the internal architecture of a whole cell without the need to dehydrate the cell, use fixatives, or potentially damaging contrast agents. This is, therefore, an ideal method for simultaneously visualizing phenotypic plasticity and the cellular location of critical macromolecules.



CAD drawing of XM2, a new soft x-ray microscope for biological and bio-energy research at the Advanced Light Source, Berkeley



Yeast cells mounted in a thin-walled capillary tube, ready for imaging using soft x-ray microscopy

An important factor in the applicability of this new technique to Genomics:GTL and Bioenergy research is the level of sample throughput that can be sustained. Typically, a complete tomographic data set can be collected from a bacterial cell in less than three minutes. In addition, upwards of twenty cells can be stacked horizontally in a glass capillary sample mounting device. A small translation of this capillary at the end of data collection quickly brings fresh cells in to the field of view, ready for imaging. In this way, information on the sub-cellular architecture of the bacterium, and the locations where critical molecular interactions take place, can be obtained with statistical significance in a short space of time.

We are currently applying soft x-ray microscopy to the study of cell development, and formation of the polar regions in the bacteria *Caulobacter crescentus*. Towards this end, we are developing technologies for labeling proteins in this bacteria for imaging by both light and x-ray microscopy. Details of these methods will be presented on the poster

Funded jointly by the NIH and the DOE, the new soft x-ray microscope was constructed at the worlds brightest soft x-ray source, the Advanced Light Source of Lawrence Berkeley National Laboratory. The new instrument was completed at the end of 2006, and is now being commissioned. The facilities are open to use by any qualified scientist. The external user program is expected to commence, in limited numbers, in the summer of 2007.

References

1. X-ray tomography of whole cells, MA Le Gros, G. McDermott, & CA Larabell. (2005) *Current Opinion in Structural Biology*, **15**, 593-600

GTL Milestone 2

Develop Methods and Concepts Needed to Achieve a Systems-Level Understanding of Microbial Cell and Community Function, Regulation, and Dynamics

Section 1

OMICS: Systems Measurements of Plants, Microbes, and Communities

54 ^{GTL}

The Virtual Institute of Microbial Stress and Survival: An Overview of the Environmental Stress Pathway Project

Carl Abulencia,^{4,14} Eric J. Alm,^{4,12} Gary Anderson,^{1,4} Edward Baidoo,^{1,4,7} Peter Benke,^{1,4,7} Sharon Borglin,^{1,4} Eoin L. Brodie,^{1,4} Romy Chakraborty,^{1,4} Swapnil Chhabra,^{1,4} Gabriela Chirica,^{2,4} Dylan Chivian,^{1,4} Michael J. Cipriano,^{1,4} M.E. Clark,^{4,9} Paramvir S. Dehal,^{1,4} Elliot C. Drury,^{4,8} Inna Dubchak,^{1,4,5} Dwayne A. Elias,^{4,8} Matthew W. Fields,^{4,9} J. Gabster,^{2,4} Sara P. Gaucher,^{3,4} Jil Geller,^{1,4} B. Giles,^{4,8} Masood Hadi,^{3,4} Terry C. Hazen,^{1,4} Qiang He,^{4,11} Zhili He,^{4,10} Christopher L. Hemme,^{4,10} E. Hendrickson,^{2,4} Kristina L. Hillesland,^{2,4} Hoi-Ying Holman,^{1,4} Katherine H. Huang,^{1,4} Y. Wayne Huang,^{1,4} C. Hwang,^{4,9} Janet Jacobsen,^{1,4} Marcin P. Joachimiak,^{1,4} Dominique C. Joyner,^{1,4} Jay D. Keasling,^{1,4,7} Keith Keller,^{1,4} Martin Keller,^{4,6} J. Leigh,^{2,4} T. Lie,^{2,4} Aindrila Mukhopadhyay,^{1,4} Richard Phan,^{1,4} Francesco Pingitore,^{1,4,7} Morgan Price,^{1,4} Alyssa M. Redding,^{1,4,7} Joseph A. Ringbauer Jr.,^{4,8} Rajat Sapra,^{3,4} Christopher W. Schadt,^{4,6} Amy Shutkin,^{1,4} Anup K. Singh,^{3,4} David A. Stahl,^{2,4} Sergey M. Stolyar,^{2,4} Yinjie Tang,^{1,4} Joy D. Van Nostrand,^{4,10} Chris B. Walker,^{2,4} Judy D. Wall,^{4,8} Eleanor Wozel,^{1,4} Zamin K. Yang,^{4,6} Huei-Che Yen,^{4,8} Grant Zane,^{4,8} Aifen Zhou,^{4,10} Jizhong Zhou,^{4,10} and **Adam P. Arkin**^{1,4,7,13*} (aparkin@lbl.gov)

¹Lawrence Berkeley National Laboratory, Berkeley, California; ²University of Washington, Seattle, Washington; ³Sandia National Laboratories, Livermore, California; ⁴Virtual Institute for Microbial Stress and Survival, <http://vimss.lbl.gov>; ⁵DOE Joint Genome Institute, Walnut Creek, California; ⁶Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee; ⁷University of California, Berkeley, California; ⁸Department of Biochemistry, University of Missouri, Columbia, Missouri; ⁹Department of Microbiology, Miami University, Oxford, Ohio; ¹⁰Institute for Environmental Genomics, University of Oklahoma, Norman, Oklahoma; ¹¹Civil and Environmental Engineering, Temple University, Philadelphia, Pennsylvania; ¹²Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts; ¹³Howard Hughes Medical Institute, Chevy Chase, Maryland; and ¹⁴Diversa, Inc., San Diego, California

Project Goals: Environmental Stress Pathway Project (ESPP) is developing computational models that describe and predict the behavior of gene regulatory networks in microbes in response to the environmental conditions found in DOE waste sites. The research takes place within the Virtual Institute for Microbial Stress and Survival (VIMSS). Based at Lawrence Berkeley National Laboratory (LBNL), VIMSS supports an integrated and multi-institutional program to understand the ability of bacteria and other microorganisms to respond to and survive external stresses. VIMSS was established in 2002 with funding from the U.S. Department of Energy Genomics:GTL Program for Rapid Deduction of Stress Response Pathways in Metal/Radionuclide Reducing Bacteria. LBNL is operated by the University of California for the U.S. Department of Energy.

The Virtual Institute of Microbial Stress and Survival (VIMSS, <http://vimss.lbl.gov>) was established in 2002 through DOE Genomics:GTL funding of the Environmental Stress Pathway Project (ESPP). The mission of VIMSS is to create a coordinated core community of scientists and a supporting experimental and computational infrastructure for the large scale comparative systems biological study of microbes and the processes they perform in different environments.

As the founding project, the central goal of ESPP is to help build this VIMSS organization and apply the resources for the deduction of regulatory pathways and systems that impact the ability of bacteria to reduce metals and radionuclides in contaminated soil. To this end, over the past four and half years the ESPP team has built capabilities in environmental chemical and microbial monitoring and perturbation; environmental microbe isolation and characterization; controlled biomass production for anaerobic microbes and co-cultures of mixed microbes; phenotypic characterization of microbes; physiological imaging of microbes; development of genetic systems for environmental microbes; transcript, protein and metabolite profiling of single microorganisms and simple co-culture; and computational analysis of these diverse data types in comparative genomic context. With these tools at hand ESPP has focused on uncovering the pathways that the sulfate reducing bacterium *Desulfovibrio vulgaris* employs to respond to environmental cues, scavenge energy from the sulfate, reduce metals and interact with other members of its community. We have learned a great deal about the evolutionary biology of these pathways and how this has impacted the organization of its genome. We have been able to compare the physiology of *D. vulgaris* to other environmental microbes to discover how regulatory strategies may be tuned to deal with different environmental niches. This work has resulted in over 73 publications to date. To aid outside researchers to access this data and other results of the VIMSS team, ESPP has built the Experimental Information and Data Repository (EIDR, <http://vimss.lbl.gov/EIDR/>) and the MicrobesOnline comparative functional genomics Workbench (<http://www.microbesonline.org/>) which also incorporate community tools so other researchers can annotate and add data.

Since its inception with ESPP, VIMSS has grown to house or formally collaborate with seven large scale projects (three of which are other Genomics:GTL projects) and informally supports several others in genome annotation, data analysis, and use of the VIMSS experimental pipeline.

55 ^{GTL}**VIMSS ESPP Functional Genomics Core: Cell Wide Analysis of Metal-Reducing Bacteria**

Aindrila Mukhopadhyay,^{1,6*} Edward Baidoo,^{1,6} Peter Benke,^{1,6} Swapnil Chhabra,^{1,6} Gabriela Chirica,^{2,6} Elliot Drury,^{3,6} Matthew Fields^{5,6} (fieldsmw@muohio.edu), Sara Gaucher,^{2,6} Masood Hadi,^{2,6} Qiang He,^{4,6} Zhili He,^{4,6} Chris Hemme,^{4,6} **Jay Keasling**^{1,6} (keasling@berkeley.edu), Francesco Pingitore,^{1,6} Alyssa Redding,^{1,6} Rajat Sapra,^{2,6} Anup Singh^{2,6} (aksingh@sandia.gov), Yinjie Tang,^{1,6} Judy Wall^{3,6} (wallj@missouri.edu), Huei-Che Yen,^{3,6} Grant Zane,^{3,6} Aifen Zhou,^{4,6} and Jizhong Zhou^{4,6} (jzhou@rccc.ou.edu)

¹Lawrence Berkeley National Laboratory, Berkeley, California; ²Sandia National Laboratories, Livermore, California; ³University of Missouri, Columbia, Missouri; ⁴University of Oklahoma, Norman, Oklahoma; ⁵Miami University, Oxford, Ohio; and ⁶Virtual Institute for Microbial Stress and Survival (VIMSS; vimss.lbl.gov)

Project Goals: Environmental Stress Pathway Project (ESPP) is developing computational models that describe and predict the behavior of gene regulatory networks in microbes in response to the environmental conditions found in DOE waste sites. The research takes place within the Virtual Institute for Microbial Stress and Survival (VIMSS). Based at Lawrence Berkeley National Laboratory (LBNL), VIMSS supports an integrated and multi-institutional program to understand the ability of bacteria and other microorganisms to respond to and survive external stresses. VIMSS was established in 2002 with funding from the U.S. Department of Energy Genomics:GTL Program for Rapid Deduction of Stress Response Pathways in Metal/Radionuclide Reducing Bacteria. LBNL is operated by the University of California for the U.S. Department of Energy.

Heavy metal and radionuclide waste poses a serious problem as a source of environmental contamination. The discovery of dissimilatory sulfate and metal ion-reducing bacteria in such ecosystems presents the opportunity to develop effective biocontainment strategies. Such protocols have the potential to be cost-effective and involve limited physical perturbation towards the environment. *Desulfovibrio vulgaris* Hildenborough is an anaerobic mesophile that belongs to the sulfate-reducing class of bacteria found ubiquitously in nature. While historically studied because of its role in bio-corrosion of oil and gas pipelines, *D. vulgaris* has the ability to reduce a wide variety of metals and radionuclides. A rigorous understanding of *D. vulgaris* physiology and its ability to survive in its environment will be critical to discern the biogeochemistry at metal contaminated sites, for bioremediation and natural attenuation for toxic metals. The availability of an annotated genomic sequence provides the foundation for such studies conducted in the functional genomics core of the VIMSS Environmental Stress Pathway Project (ESPP).

Utilizing our optimized pipeline for generating biomass for functional genomics studies, we conducted transcriptomics analysis to measure *D. vulgaris* response to mutations in important global regulators such as Fur, PerR and Zur. The response of these mutants to environmental stresses was also examined. Additionally we continued to collect data in *Shewanella oneidensis* and *Geobacter metallireducens* for comparative genomics. More importantly we extended transcriptomics analysis to examine alternate *D. vulgaris* physiological states such as in biofilms and growth in syntrophic co-culture with *Methanococcus maripaludis*. In order to address cell wide responses beyond the initial changes at the mRNA level, a quantitative proteomics workflow using the iTRAQ peptide tagging

strategy was optimized to provide proteins change measurements. This method provides data for 30% of the proteome and covers all known protein functional categories. Cell wide proteomics and transcriptomics measurements are now being used to develop models for several of these environmentally relevant stresses such as oxygen exposure, oxidative stress, as well as the altered physiology of biofilms and co-culture. Continued studies to map cell wide responses have also emphasized the importance of changes that require orthologous measurements – such as monitoring post translational modifications and protein-protein interactions. With this in mind a novel protocol to monitor redox state of the proteins has been developed as well exogenous protocols to study protein-protein interactions. Measurements at the metabolite and metabolic pathways level are necessary to complete our cell wide studies. Metabolite extraction and detection for several hundred metabolites can now be conducted for these non-model organisms using high resolution separation and mass spectroscopic methods. Flux analysis methods were established using ^{13}C -Lactate grown *D. vulgaris* and *S. oneidensis*. The use of a novel high resolution FTICR-MS method for flux analysis led to results that allowed us to re-evaluate the genome annotation for the central metabolic pathways in *D. vulgaris*. The integration of these methods enables a comprehensive understanding of *D. vulgaris* physiology.

56 ^{GTL}

Response of *Desulfovibrio vulgaris* Hildenborough to Acid pH

H.-C. Yen,^{1,7*} T. C. Hazen,^{2,7} Z. Yang,^{3,7} J. Zhou,^{4,7} K. H. Huang,^{5,7} E. J. Alm,^{5,7} **A. P. Arkin**^{6,7} (aparkin@lbl.gov), and J. D. Wall^{1,7}

¹Department of Biochemistry, University of Missouri, Columbia, Missouri; ²Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California; ³Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee; ⁴Department of Botany and Microbiology, University of Oklahoma, Norman, Oklahoma; ⁵Massachusetts Institute of Technology, Cambridge, Massachusetts; ⁶Department of Bioengineering, University of California, Berkeley, California; ⁷Virtual Institute for Microbial Stress and Survival, <http://vimss.lbl.gov/>

Project Goals: The DOE oversees 350 cleanup projects involving soil contaminated with metals or radionuclides. The life-cycle cost of these projects is \$220 billion over 70 years and could be \$300 billion without breakthroughs. One breakthrough approach may be to exploit bacteria that can immobilize and detoxify metals in soil via reduction to less soluble and less toxic forms. This occurs naturally and be stimulated in situ. A thorough understanding of the biogeochemistry, especially stress responses in metal/radionuclide bacteria, will enable prediction of natural attenuation and new strategies for remediation that could save DOE billions in cleanup, risk assessment, and environmental stewardship. To achieve this understanding we will study three such organisms that occupy different niches at these sites by developing validated culture conditions similar to site conditions, and then using functional genomic and comparative genomic analysis to deduce the molecular basis of the stress responses that affect metal reduction efficiency. Models of the physiological processes will be made of accuracy suitable for computation development of field protocols for stimulating metal reduction. In addition, comparison of pathways among niches will give insight into how the bacteria adapt their pathways to different conditions.

The anaerobic, sulfate-reducing bacterium *Desulfovibrio vulgaris* Hildenborough has been suggested to be useful for environmental bioremediation applications. Among the constraints likely to be encountered, that effect the growth of a bacterial community in contaminated sites, are low pH con-

ditions. Growth of *D. vulgaris* on acid pH media was studied. It was evident that this bacterium was able to grow at pHs as low as 5 in batch cultures; however, the lag phase was prolonged and the final protein yield was proportionally lower with the decrease in pH. In medium with lactate as the carbon and reductant source and sulfate as the terminal electron acceptor, the final protein yields dropped to 50% or 29% of the control (pH 7) when the initial medium pH was pH 5.5 or pH 5.0, respectively. The average lag phase for initial pH 5.5 medium was 120 hours versus pH 7 which generally had two hours or less.

This bacterium incompletely oxidizes organic acids with the production of acetate. Thus at low pH, growth is limited by the accumulation of acetate in the medium that acts to shuttle protons across the membrane thereby dissipating the proton gradient and acidifying the cytoplasm. The production of sulfide gas through sulfate reduction consumes protons thereby countering the acid conditions. Finally deamination or decarboxylation of certain amino acids, e.g. lysine, arginine, tryptophan, and isoleucine, may gradually alkalize the acid medium. When the medium pH is high enough (~6.5) for lactate oxidation and sulfate reduction, cell growth follows.

Transcriptional profiling was performed for *D. vulgaris* exposed to pH 5.5 and 6.2. Typical acid shock responses for Gram-negative bacteria were not observed. Instead, from the transcripts with the highest differential expression, it was inferred that significant cellular damage had resulted because chaperone genes including heat shock genes and repair genes for both proteins and nucleic acids were greatly induced.

57 ^{GTL}

Global Gene Regulation in *Desulfovibrio vulgaris* Hildenborough

Aifen Zhou,^{1,5*} Zhili He,^{1,5} Chris Hemme,^{1,5} Aindrila Mukhopadhyay,^{2,5} Jay Keasling,^{2,5} **Adam P. Arkin**^{2,5} (aparkin@lbl.gov), Terry C. Hazen,^{2,5} Judy D. Wall,^{4,5} and Jizhong Zhou^{1,5}

¹Institute for Environmental Genomics, Department of Botany and Microbiology, University of Oklahoma, Norman, Oklahoma; ²Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California; ³Departments of Biochemistry and Molecular Microbiology & Immunology, University of Missouri, Columbia, Missouri; ⁴Earth Science Division, Lawrence Berkeley National Laboratory, Berkeley, California; and ⁵Virtual Institute for Microbial Stress and Survival, <http://vimss.lbl.gov>

Project Goals: Environmental Stress Pathway Project (ESPP) is developing computational models that describe and predict the behavior of gene regulatory networks in microbes in response to the environmental conditions found in DOE waste sites. The research takes place within the Virtual Institute for Microbial Stress and Survival (VIMSS). Based at Lawrence Berkeley National Laboratory (LBNL), VIMSS supports an integrated and multi-institutional program to understand the ability of bacteria and other microorganisms to respond to and survive external stresses. VIMSS was established in 2002 with funding from the U.S. Department of Energy Genomics:GTL Program for Rapid Deduction of Stress Response Pathways in Metal/Radionuclide Reducing Bacteria. LBNL is operated by the University of California for the U.S. Department of Energy.

Desulfovibrio vulgaris Hildenborough (DvH) is an obligate anaerobe and has been used as a model organism for studying the energy metabolism of sulfate-reducing bacterium (SRB). However, experimental data about the transcriptional regulatory networks which are essential for understanding the cellular processes are very limited. One of the central goals of ESPP is to link laboratory measurements of stress responses and metabolism to activities of microbes in the field. Thus, towards this goal, we have performed laboratory study of gene expression and regulations in *D. vulgaris* in responses to oxidative stress and the importance of key global regulatory genes in stress responses. Several predicted global regulators are investigated via mutant characterization, transcriptomic assay and their *in vivo* gene regulations using ChIP-chip assay.

CRP/FNR. CRP/FNR regulators are DNA binding proteins function as positive transcription factors. There are four CRP/FNR homologues in the DvH genome (DVU2547, DVU0379, DVU3111, DVU2097). Evidence from other bacteria demonstrated that CRP/FNR regulators function in response to a broad spectrum of intracellular and exogenous signals such as oxidative and nitrosative stress, nitric oxide, carbon monoxide or temperature. Microarray data from DvH shows that their transcript levels are altered in response to nitrate, nitrite, heat shock, and oxygen stresses. To determine the function to the DvH CRP/FNR, knockout mutants for all four CRP/FNR proteins were generated. The mutants will be characterized using various electron donors and acceptors, different stressors, and transcriptomic analysis. To study the global gene regulation by CRP/FNR, recombinant proteins for all four CRP/FNR were obtained and polyclonal antibodies were generated. Immunoprecipitated DNA-protein complexes with specific CRP/FNR polyclonal antibodies will be hybridized to the DvH PCR-amplicon promoter array. And the CRP/FNR binding motif can be identified by computational and experimental approaches.

H₂O₂ stress response. Oxidative stress is one of the most common environmental stressors. Evidences show that DvH cells are aero-tolerant although they are strict anaerobe. But little is known about molecular mechanisms of oxidative stress responses. DvH is one of the few microorganisms that contain both defense systems which are typical for the aerobic (Sod and Kat) and the anaerobic (Rub, Rbr, Rbo etc.) microbes, but their roles remain elusive. In this study, DvH cells were stressed with two different concentrations of H₂O₂ (1 mM, 4 mM) and 5 time-points (30, 60, 120, 240 and 480 min) were used for the transcriptomic analysis. Microarray data demonstrated that higher concentration of H₂O₂ had broader effect on gene expression. The time-points with the greatest gene expression changes are 120 min (485 up and 527 down) and 240 min (750 up and 753 down) for 1 mM and 4 mM of H₂O₂ respectively. Rdl, Rbr2 were up-regulated, which suggest that these two proteins, rather than Rub-Rbo & Rbr suggested by Coulter's *in vitro* experiment data, may play major roles in H₂O₂ stress. Genes in the predicted PerR and FUR regulon were also up-regulated. Some interesting candidates such as DVU3269 (a hybrid histidine kinase (HK)), DVU3136 (a nitroreductase family protein) etc. were significantly up-regulated. The function of the putative candidates in H₂O₂ stress response are going to be confirmed by other approaches such as knockout mutant analysis.

FUR, PerR and ZUR. FUR, PerR and ZUR are Fur Paralogs in the DvH genome. Microarray data show that they are involved in iron acquisition, acid shock response and oxidative stress etc. The functional analysis of these three global transcription regulators is in progress.

58 [—]GTL***Desulfovibrio vulgaris* Responses to Hexavalent Chromium at the Community, Population, and Cellular Levels**

A. Klonowska,^{1*} Z. He,^{2,6} Q. He,^{3,6} M.E. Clark,^{1,6} S.B. Thieman,¹ T.C. Hazen,^{4,6} E.L. Brodie,^{4,6} R. Chakraborty,^{4,6} E.J. Alm,^{4,6} B. Giles,^{5,6} H.-Y. Holman,^{4,6} **A.P. Arkin**^{4,6} (aparkin@lbl.gov), J.D. Wall,^{5,6} J. Zhou,^{2,6} and M.W. Fields^{1,6}

¹Department of Microbiology, Miami University, Oxford, Ohio; ²Institute for Environmental Genomics, University of Oklahoma, Norman, Oklahoma; ³Civil and Environmental Engineering, Temple University, Philadelphia, Pennsylvania; ⁴Lawrence Berkeley National Laboratory, Berkeley, California; ⁵Department of Biochemistry, University of Missouri, Columbia, Missouri; ⁶Virtual Institute for Microbial Stress and Survival, <http://vimss.lbl.gov/>

Project Goals: Environmental Stress Pathway Project (ESPP) is developing computational models that describe and predict the behavior of gene regulatory networks in microbes in response to the environmental conditions found in DOE waste sites. The research takes place within the Virtual Institute for Microbial Stress and Survival (VIMSS). Based at Lawrence Berkeley National Laboratory (LBNL), VIMSS supports an integrated and multi-institutional program to understand the ability of bacteria and other microorganisms to respond to and survive external stresses. VIMSS was established in 2002 with funding from the U.S. Department of Energy Genomics:GTL Program for Rapid Deduction of Stress Response Pathways in Metal/Radionuclide Reducing Bacteria. LBNL is operated by the University of California for the U.S. Department of Energy.

Desulfovibrio vulgaris is an anaerobic sulfate-reducing bacterium (SRB) able to reduce toxic heavy metals such as chromium and uranium, and *D. vulgaris* represents a useful SRB model for the bioremediation of heavy metal contamination. In order to correlate cellular responses to ecosystem changes with respect to heavy metal bioremediation, measurements are needed at different ecological scale. The VIMSS group has developed a multi-institutional collaboration to better understand the relationships between field site observations, cellular responses, and biochemical processes. The current work describes the observation of *Desulfovibrio* populations at a Cr-contaminated field site within the context of possible mechanisms for Cr bio-reduction. In a field scale trial of polylactate stimulated chromate bioreduction at Hanford, *Desulfovibrio* were rapidly enriched and remained at elevated densities for at least one year. Importantly this enrichment of *Desulfovibrio* corresponded to decreased chromate concentrations which remained below background concentrations during that time. Although much work has focused on Cr and U reduction via individual enzymes, less is known about the cellular response to heavy metal stress in *Desulfovibrio* species. Cells were cultivated in a defined medium with lactate and sulfate, and a sub-lethal concentration of Cr(VI) was added at mid-exponential phase growth. Based upon microarray data, the FMN-dependent nitroreductase might reduce Cr(VI) directly or reduce a Cr-complex. The FMN reductase could synthesize FMNH₂ and the NADP dehydrogenase might be used to regenerate NADPH₂. The *chrAB* genes on the megaplasmid most likely play a key role in Cr(III) efflux based upon microarray data and growth data. Additional toxicological effects could be occurring once the Cr(III) is produced via protein denaturation in the cytoplasm, periplasm, and outer cell proper. Growth data with washed cells showed an increased sensitivity in the presence of Cr(VI). When exponential-phase cells were washed to remove hydrogen sulfide carry-over and inoculated into fresh medium with different levels of Cr(VI), lag time increased as the levels of Cr increased. Cells lagged approximately 5, 40, and 55 h in the pres-

ence of 20, 50, and 100 μM Cr, respectively. When cells were transferred to 50 μM Cr, Cr(VI) levels declined within 2 h and lactate was consumed, but sulfate did not decline until growth was initiated approximately 40 h later. Lactate continued to be consumed at a slow rate during the lag but sulfate levels remained unchanged. The results indicated that lactate oxidation was decoupled from sulfate reduction in the presence of Cr(VI). These data coincide with the working model that electrons are re-routed away from sulfate reduction and used to reduce Cr(VI). The down-expression of sulfate permease coincided with these results. The results also suggested that cells responded to the presence of Cr even after reduction based upon growth responses, FTIR analyses on cellular macromolecules, and expressed genes detected with transcriptomics.

59 [—]GTL

Energy Conserving Hydrogenases Drive Syntrophic Growth of *Desulfovibrio vulgaris* and *Methanococcus maripaludis*

C.B. Walker,^{1,6*} Z.K. Yang,^{2,6} Z. He,^{3,6} S.S. Stolyar,^{1,6} J. Jacobsen,^{4,6} J.A. Ringbauer Jr.,^{5,6} J.D. Wall,^{5,6} J. Zhou,^{3,6} **A. P. Arkin**^{4,6} (aparkin@lbl.gov), and D.A. Stahl^{1,6}

¹University of Washington, Seattle, Washington; ²Oak Ridge National Laboratory, Oak Ridge, Tennessee; ³University of Oklahoma, Norman, Oklahoma; ⁴Lawrence Berkeley National Laboratory, Berkeley, California; ⁵University of Missouri, Columbia, Missouri; and ⁶Virtual Institute for Microbial Stress and Survival, <http://vimss.lbl.gov>

Project Goals: Environmental Stress Pathway Project (ESPP) is developing computational models that describe and predict the behavior of gene regulatory networks in microbes in response to the environmental conditions found in DOE waste sites. The research takes place within the Virtual Institute for Microbial Stress and Survival (VIMSS). Based at Lawrence Berkeley National Laboratory (LBNL), VIMSS supports an integrated and multi-institutional program to understand the ability of bacteria and other microorganisms to respond to and survive external stresses. VIMSS was established in 2002 with funding from the U.S. Department of Energy Genomics:GTL Program for Rapid Deduction of Stress Response Pathways in Metal/Radionuclide Reducing Bacteria. LBNL is operated by the University of California for the U.S. Department of Energy.

In the absence of an electron acceptor, many *Desulfovibrio* species grow on non-fermentable substrates via syntrophic association with hydrogen consuming methanogens. Building upon the ongoing Virtual Institute for Microbial Stress and Survival (VIMSS) investigation into the response of *Desulfovibrio vulgaris* Hildenborough to environmental stressors found in contaminated DOE sites, the Environmental Stress Pathway Project's (ESPP) Applied Environmental Core (AEC) developed and maintained a stable syntrophic consortium. *Desulfovibrio vulgaris* Hildenborough and *Methanococcus maripaludis* LL were continuously grown in a chemostat on minimal media amended with lactate but lacking electron acceptor. Replicated whole genome transcriptional analyses by the ESPP Functional Genomics Core (FGC) and the Computational Core (CC) identified 169 and 254 genes that were significantly up-regulated or down-regulated, respectively, relative to a sulfate-limited monoculture growing at the same generation time. The majority of up-regulated genes were associated with energy production/conservation, signal transduction mechanisms, and amino acid transport/metabolism. A number of the down-regulated genes were associated with signal transduc-

tion mechanisms, inorganic ion transport/metabolism and amino acid transport and metabolism. Among those genes most highly up-regulated were a suite of hydrogenases including the putative carbon-monoxide induced hydrogenase (Coo, DVU2286 - 93). Coo is a multi-subunit membrane-bound complex with high similarity to an energy conserving protein in *Rhodospirillum rubrum*. In order to further elucidate the possible role energy conserving hydrogenases play in syntrophic growth, we examined transposon mutants generated by the FGC of both the Coo hydrogenase (*cooL*) and a structurally related homolog, Ech (energy conserving hydrogenase, *echA*, DVU0429-34). Both mutants grew to the same cell density on lactate-sulfate, although the *cooL* mutant grew significantly slower. When grown in coculture with *M. maripaludis* without any sulfate, the *cooL* mutant grew significantly slower and to approximately 25% yield, while the *echA* mutant showed a less pronounced difference in growth rate and yield (approximately 80%). Together these data suggest an important role for the Coo hydrogenase in energy conservation of *D. vulgaris* Hildenborough during syntrophic growth, possibly through proton translocation, although the exact physiological mechanism remains to be elucidated. Continued collaborative work by the VIMSS three ESPP core groups should provide a more complete mechanistic understanding of sulfate-reduction and syntrophic cooperation between microbes.

60 ^{GTL}

A Large Number of Hypothetical Proteins are Differentially Expressed during Stress in *Desulfovibrio vulgaris*

Elliot C. Drury,^{1,5*} Alyssa M. Redding,^{2,5} Aindrila Mukhopadhyay,^{2,5} Katherine H. Huang,^{3,5} Terry C. Hazen,^{2,5} **Adam P. Arkin**^{2,4,5} (aparkin@lbl.gov), Judy D. Wall,^{1,5} and Dwayne A. Elias^{1,5*}

¹Department of Biochemistry, University of Missouri, Columbia, Missouri; ²Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California; ³Massachusetts Institute of Technology, Cambridge, Massachusetts; ⁴Department of Bioengineering, University of California, Berkeley, California; and ⁵Virtual Institute for Microbial Stress and Survival, <http://vimss.lbl.gov>

Project Goals: Environmental Stress Pathway Project (ESPP) is developing computational models that describe and predict the behavior of gene regulatory networks in microbes in response to the environmental conditions found in DOE waste sites. The research takes place within the Virtual Institute for Microbial Stress and Survival (VIMSS). Based at Lawrence Berkeley National Laboratory (LBNL), VIMSS supports an integrated and multi-institutional program to understand the ability of bacteria and other microorganisms to respond to and survive external stresses. VIMSS was established in 2002 with funding from the U.S. Department of Energy Genomics:GTL Program for Rapid Deduction of Stress Response Pathways in Metal/Radionuclide Reducing Bacteria. LBNL is operated by the University of California for the U.S. Department of Energy.

Hypothetical and conserved hypothetical proteins consistently make up 30% or more of sequenced bacterial genomes, with few reports confirming their expression at either the rRNA or protein level. It is likely that many of these proteins serve significant functions ranging from regulation to presently unknown steps in carbon or electron flux pathways. Hence, the elucidation of their function(s) is highly relevant to the Virtual Institute Microbial Stress and Survival/Environmental Stress Pathway Project mission. We have compiled expression profiles for the expected 1167 hypothetical

and conserved hypothetical proteins in *D. vulgaris* from data obtained in the VIMSS/ESPP project over 10 environmental stresses, along with corresponding transcriptomic and MS-based iTRAC proteomic datasets from controlled cultures.

The genes were divided into two groups; those in polycistronic operons and those that are monocistronic. For operonic and monocistronic genes respectively, we observed 37 and 46 genes that are not expressed, 36 and 12 that show no stress-related response but are expressed at high rates, 0 and 173 that show no stress-related response but are expressed at low rates, 445 and 199 with significant response in two or more stresses, and 104 and 123 that only showed significant differential expression in one stress. While the number of transcription studies outweighs protein analysis, the abundance values indicating differential expression at the protein level were highly consistent with the microarray results when data were available.

Therefore, we are presently able to confirm, at both the mRNA and protein level, the expression of 253 hypothetical and conserved hypothetical proteins and show no evidence for 83 genes encoding a protein. Those that are expressed should be re-annotated to “expressed protein” while the remainder should be described as “non-coding gene”. Of those that did show expression, elucidating function without a stress-related expression pattern is difficult, particularly for monocistronic genes in this category. The proteins that showed differential expression in response to one or more stresses are theoretically easier to deduce a putative function, especially for those in operons, and these assignments have been completed. Finally, the validity of such putative assignments can only be ascertained by interruption or deletion of the gene with further analysis. We are in the process of testing six mutants that have been isolated from a random transposon library. As the library continues to be sequenced, we will test the interrupted hypothetical and conserved hypothetical proteins either to assign a function or to confirm the putative assignments. By confidently confirming the function of these proteins and the effect of their removal in *D. vulgaris*, there will be a more thorough understanding of the mechanisms by which this bacterium survives stresses likely experienced at DOE contaminated sites.

61 ^{GT}L

Phenotypic Correlations in *Desulfovibrio*

K.L. Hillesland,* C.B. Walker, and D.A. Stahl

University of Washington, Seattle, Washington and Virtual Institute for Microbial Stress and Survival, <http://vimss.lbl.gov>

Project Goals: Environmental Stress Pathway Project (ESPP) is developing computational models that describe and predict the behavior of gene regulatory networks in microbes in response to the environmental conditions found in DOE waste sites. The research takes place within the Virtual Institute for Microbial Stress and Survival (VIMSS). Based at Lawrence Berkeley National Laboratory (LBNL), VIMSS supports an integrated and multi-institutional program to understand the ability of bacteria and other microorganisms to respond to and survive external stresses. VIMSS was established in 2002 with funding from the U.S. Department of Energy Genomics:GTL Program for Rapid Deduction of Stress Response Pathways in Metal/Radionuclide Reducing Bacteria. LBNL is operated by the University of California for the U.S. Department of Energy.

An ESPP goal is to distinguish genetic pathways that evolved as an adaptation to stressors from those that arose simply through inheritance from ancestral species. To do so, it is necessary to characterize genetic relationships that are conserved across broader taxonomic groups. Phenotypic correlations that are evident within a genus are likely to result from conserved pleiotropic relationships among traits (i.e. both traits share at least some genes in common). Such relationships may accelerate or limit evolutionary adaptation depending on whether they are positive or negative, respectively. Thus, identifying pleiotropy may lead to new insights into the evolutionary trajectories of a taxon as well as its physiology. To expand our knowledge of the physiology and evolution of the genus *Desulfovibrio*, we measured several features of growth of 14 strains in the presence of different electron donors (lactate or pyruvate) and acceptors (sulfate, no electron acceptor, or in coculture with the hydrogenotrophic methanogen, *Methanococcus maripaludis*). We observed a strong positive correlation between growth rate on lactate and pyruvate when sulfate was the electron acceptor ($r = 0.79$, $p = 0.0007$), but not when *M. maripaludis* was the surrogate electron acceptor. However, there was a positive correlation between the biomass achieved on lactate versus pyruvate with *M. maripaludis* as the electron acceptor ($r = 0.8$, $p = 0.0033$). It is possible that the growth rate of *Desulfovibrio* on lactate and pyruvate was also correlated, but variability in growth kinetics of *M. maripaludis* obscured our ability to detect it. The relationships among growth phenotypes on different electron acceptors were not consistent among electron donors. Growth rate on pyruvate with *M. maripaludis* was positively correlated with growth rate on pyruvate and sulfate ($r = 0.8$, $p = 0.0028$) and lactate and sulfate ($r = 0.75$, $p = 0.0076$). These correlations were not evident when lactate was the electron donor. In fact, with lactate as the electron donor, growth rate with sulfate as the electron acceptor was negatively correlated with the level of biomass achieved from growth with *M. maripaludis* as the electron acceptor ($r = -0.77$, $p = 0.006$). In addition to these correlations indicating potential pleiotropic relationships, individual strains showed unusual characteristics. For example, among this *Desulfovibrio* study set, only *D. vulgaris* Llanelly was unable to ferment pyruvate. Although this strain was incapable of syntrophic growth on lactate with *M. maripaludis*, syntrophic growth was possible with pyruvate as the electron donor. Together, these results suggest that some genes involved in mechanisms of energy acquisition are conserved among strains that have the capacity for syntrophic growth. The ecological consequences of these relationships between traits will be explored in experiments examining the relative fitness among *Desulfovibrio* growing under different conditions of electron donor and electron acceptor availability.

62 ^{GTL}

Nitrate Stress Response in *Desulfovibrio vulgaris* Hildenborough: Whole-Genome Transcriptomics and Proteomics Analyses

Qiang He,^{1,8*} Zhili He,^{2,8} Wenqiong Chen,³ Zamin Yang,^{4,8} Eric J. Alm,^{5,8} Katherine H. Huang,^{5,8} Huei-Che Yen,^{6,8} Dominique C. Joyner,^{7,8} Martin Keller,^{3,4,8} **Adam P. Arkin**^{5,8} (aparkin@lbl.gov), Terry C. Hazen,^{7,8} Judy D. Wall,^{6,8} and Jizhong Zhou^{2,8}

¹Department of Civil & Environmental Engineering, Temple University, Philadelphia, Pennsylvania;

²Institute for Environmental Genomics, Department of Botany and Microbiology, University of Oklahoma, Norman, Oklahoma; ³Diversa Corporation, San Diego, California; ⁴Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee; ⁵Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California; ⁶Departments of Biochemistry and Molecular Microbiology & Immunology, University of Missouri, Columbia, Missouri; ⁷Earth Science Division, Lawrence Berkeley National Laboratory, Berkeley, California; and ⁸Virtual Institute for Microbial Stress and Survival, <http://vimss.lbl.gov>

Project Goals: Environmental Stress Pathway Project (ESPP) is developing computational models that describe and predict the behavior of gene regulatory networks in microbes in response to the environmental conditions found in DOE waste sites. The research takes place within the Virtual Institute for Microbial Stress and Survival (VIMSS). Based at Lawrence Berkeley National Laboratory (LBNL), VIMSS supports an integrated and multi-institutional program to understand the ability of bacteria and other microorganisms to respond to and survive external stresses. VIMSS was established in 2002 with funding from the U.S. Department of Energy Genomics:GTL Program for Rapid Deduction of Stress Response Pathways in Metal/Radionuclide Reducing Bacteria. LBNL is operated by the University of California for the U.S. Department of Energy.

Sulfate reducing bacteria (SRB) are of interest for bioremediation with their ability to reduce and immobilize heavy metals. Nitrate, a common co-contaminant in DOE sites, is suggested to inhibit SRB via nitrite. Previous results indicate that nitrite is indeed inhibitory to the growth of *Desulfovibrio vulgaris*. However, growth inhibition by nitrate alone was also observed. One of the central goals of ESPP is to link laboratory measurements of stress responses and metabolism to activities of microbes in the field. Thus to this end, we have performed laboratory study of expression in *D. vulgaris* in responses to nitrate stress. In this study, growth and expression responses to various concentrations of nitrate were investigated using the Omnilog phenotype arrays and whole-genome DNA microarrays. Changes in the proteome were examined with 3D-LC followed by MS-MS analysis.

Microarray analysis found 5, 50, 115, and 149 genes significantly up-regulated and 36, 113, 205, and 149 down-regulated at 30, 60, 120, and 240 min, respectively. Both transcriptomic and proteomic profiles shared little similarities with those of salt stress, indicating a specific inhibitory mechanism beyond osmotic stress. Many of the genes (~50% at certain time points) with altered expression level were of unknown functions; however, the increasing number of ribosomal protein genes down-regulated with time could provide a direct explanation to the growth inhibition effect of nitrate. Further, several lines of evidence suggested that the down-regulation of genes coding the ribosomal proteins could be the result of the changes in the energy flow upon nitrate exposure: 1) The down-regulation of genes for the ATPase subunits indicated reduced level of energy generation; 2) the up-regula-

tion of phage shock protein genes (*pspA* and *pspC*) might indicate a reduced proton motive force; although damages to the cell envelope could also contribute to this outcome; 3) the gene for the hybrid cluster protein, a redox protein with roles in nitrogen metabolism, was highly up-regulated 120 and 240 min following nitrate stress at both transcriptomic and proteomic level, suggesting that nitrate was being actively reduced, shifting reducing equivalents away from normal energy production; 4) genes in the methionine biosynthesis pathway were among the most highly up-regulated genes throughout the experiment, potentially providing a convenient mechanism for the simultaneous disposal of excess sulfur (from sulfate reduction) and nitrogen (from nitrate reduction); 5) One gene cluster consistently among the most up-regulated genes consisted of genes encoding two TRAP dicarboxylate family transporters, a formate acetyltransferase, and a pyruvate formate-lyase activating enzyme, which might be regulated to provide an increased carbon flow to keep pace with demand from amino acids biosynthesis. These observations indicated that the growth inhibition effect of nitrate might be due to energy limitation.

Similar to the observations made during salt stress, the glycine/betaine transporter gene was among genes highly up-regulated, suggesting that NaNO_3 also constituted osmotic stress which was relieved by the mechanism of osmoprotectant accumulation. Osmoprotectant accumulation as the major resistance mechanism was further validated by the partial relief of growth inhibition by glycine betaine. It is also noted that, similar to nitrite stress, the ferric iron transporter genes were up-regulated during nitrate stress, suggesting an increased demand for iron. Unlike nitrite stress, however, no other genes in the Fur regulon were co-regulated during nitrate stress, pointing to a yet-to-known regulatory signal.

In conclusion, excess NaNO_3 resulted in both osmotic stress and nitrate stress. *D. vulgaris* shifted nitrogen metabolism and energy production in response to nitrate stress. Resistance to osmotic stress was achieved primarily by the transport of osmoprotectant.

63 ^{GTL}

Redox Proteomics in *Desulfovibrio vulgaris* Hildenborough: Search for Proteins that Mediate Stress Response via Post-Translational Modification of the Cys Residues

Rajat Sapra,^{1,2*} Sara Gaucher,^{1,2} Gabriela Chirica,² Carrie Kozina,² George Buffleben,² Richard Phan,^{1,3} Dominique Joyner,^{1,3} Terry C. Hazen,^{1,3} **Adam P. Arkin**^{1,3} (aparkin@lbl.gov), and Anup K. Singh^{1,2}

¹Virtual Institute for Microbial Stress and Survival, <http://vimss.lbl.gov>; ²Sandia National Laboratories, Livermore, California; and ³Lawrence Berkeley National Laboratory, Berkeley, California

Project Goals: Environmental Stress Pathway Project (ESPP) is developing computational models that describe and predict the behavior of gene regulatory networks in microbes in response to the environmental conditions found in DOE waste sites. The research takes place within the Virtual Institute for Microbial Stress and Survival (VIMSS). Based at Lawrence Berkeley National Laboratory (LBNL), VIMSS supports an integrated and multi-institutional program to understand the ability of bacteria and other microorganisms to respond to and survive

external stresses. VIMSS was established in 2002 with funding from the U.S. Department of Energy Genomics:GTL Program for Rapid Deduction of Stress Response Pathways in Metal/Radionuclide Reducing Bacteria. LBNL is operated by the University of California for the U.S. Department of Energy.

Desulfovibrio vulgaris Hildenborough (*D. vulgaris*) is a sulfate reducing bacterium that grows in the absence of oxygen in a reducing environment. From a physiological as well as ecological perspective, anaerobic bacteria have to redox stress in the form of reactive oxygen species (ROS) produced when the cells are exposed to molecular oxygen or chemicals like hydrogen peroxide. To counteract the deleterious effects of ROS, many anaerobic organisms have developed defense systems similar to those found in aerobes. In *D. vulgaris* enzymes that detoxify ROS include superoxide reductase (Sor) which can reduce superoxide to water and rubrerythrin (Rbr), which can reduce H_2O_2 to water without the regeneration of oxygen, a feature that is important for oxygen detoxification in anaerobe. Another mechanism which anaerobic bacteria have evolved to overcome redox stress is using post translational modifications of proteins, specifically the modifications of Cysteine (Cys) residues in proteins, to minimize the deleterious effects of the ROS and redox associated stress.

Cys is one of the most rarely used amino acids in the proteins of most organisms studied so far. Cys containing proteins are key in maintain the reducing environment of the cytosol and alleviating redox stress due to the presence of ROS. Furthermore, Cys residues in redox active proteins mediate redox reactions where transfer of electrons proceeds via thiol-disulfide exchange reactions. Importantly, all of these activities of Cys containing cytosolic enzymes usually depend on the preservation of the reduced state of the cysteine residue(s) involved. These redox active thiol groups of Cys residues can thus be post-translationally modified to form intra- and inter-molecular disulfide bonds or oxidized to sulfinic, sulfenic acid in response to redox stress.

We are investigating the mechanisms by which *D. vulgaris* proteins are post-translationally modified to counter redox stress induced by oxygen and Cr(VI) reduction using a combination of proteomics techniques that leverage the interdisciplinary cores of the VIMSS. *D. vulgaris* cells are grown in the environmental microbiology lab (Hazen Lab, Lawrence Berkeley National Labs). The protein samples generated are modified to block all free Cys residues with N-ethylmaleimide to reduce the complexity of the sample and to specifically target the modified Cys residues. The samples thus generated are surveyed using a combination of 2D-DIGE and Isotope Coded Affinity Tag (ICAT) proteomics (Singh Lab, Sandia National Labs) to identify post translational modifications and, finally, ITRAQ proteomics is used to identify the relative abundance of the proteins (Keasling Lab, Lawrence Berkeley National Labs). The stress response at the proteome level is compared to the transcriptomics data generated from the same samples by the bioinformatics core (Arkin Lab, Lawrence Berkeley National Labs) to develop a comprehensive picture of stress response and stress-induced post translational modifications of proteins in *D. vulgaris*. Using this approach, we have identified more than 50 proteins and have mapped the Cys residues in the corresponding proteins that have function associated with redox stress in *D. vulgaris*.

64 ^{GTL}**A Survey of Protein Post-Translational Modifications Found in the Sulfate-Reducing Bacterium *Desulfovibrio vulgaris* Hildenborough**

Sara P. Gaucher,^{1,5*} Alyssa M. Redding,^{2,5} Gabriela S. Chirica,¹ Rajat Sapra,^{1,5} George M. Buffleben,¹ Carrie Kozina,¹ Aindrila Mukhopadhyay,^{2,5} Dominique C. Joyner,^{2,5} Jay D. Keasling,^{2,5} Terry C. Hazen,^{2,5} **Adam P. Arkin**^{2,5} (aparkin@lbl.gov), David A. Stahl,^{4,5} Judy D. Wall,^{3,5} and Anup K. Singh^{1,5}

¹Sandia National Laboratories, Livermore, California; ²Lawrence Berkeley National Laboratory, Berkeley, California; ³University of Missouri, Columbia, Missouri; ⁴University of Washington, Seattle, Washington; and ⁵Virtual Institute for Microbial Stress and Survival, <http://vimss.lbl.gov>

Project Goals: Environmental Stress Pathway Project (ESPP) is developing computational models that describe and predict the behavior of gene regulatory networks in microbes in response to the environmental conditions found in DOE waste sites. The research takes place within the Virtual Institute for Microbial Stress and Survival (VIMSS). Based at Lawrence Berkeley National Laboratory (LBNL), VIMSS supports an integrated and multi-institutional program to understand the ability of bacteria and other microorganisms to respond to and survive external stresses. VIMSS was established in 2002 with funding from the U.S. Department of Energy Genomics:GTL Program for Rapid Deduction of Stress Response Pathways in Metal/Radionuclide Reducing Bacteria. LBNL is operated by the University of California for the U.S. Department of Energy.

Sulfate reducing bacteria (SRB), found widely in nature, use sulfate as the terminal electron acceptor in their respiratory cycle, leading to the production of hydrogen sulfide. These bacteria have both ecological and economic importance. SRB play a role in various biogeochemical cycles including the sulfur and carbon cycles. They have a negative economic impact on the oil industry, where their metabolism causes corrosion and clogging of machinery, and fouling of oil wells. However, they have also been shown to reduce and/or immobilize toxic water-soluble metals such as copper (II), chromium (IV) and uranium (VI), and thus are candidates for bioremediation applications.

Desulfovibrio vulgaris Hildenborough (DvH) is a member of the most well studied genus of SRBs. A goal of the Environmental Stress Pathway Project (ESPP) in the Virtual Institute for Microbial Stress and Survival (VIMSS) is to understand the regulatory networks in DvH for applications to bioremediation. One aspect of this is the elucidation of protein post-translational modifications (PTMs) in DvH.

PTMs play various roles in the cell. Some modifications play a role in protein structure, such as lipid anchors or some disulfide bonds. Others are directly involved in regulation of protein function such as phosphorylation and glycosylation. Still others arise through cellular damage such as irreversible oxidation events. Whatever the role these PTMs play, they must be characterized at the protein level because they are not directly coded for in the genome. Furthermore, DvH may be particularly likely to use PTMs as a regulatory mechanism: Evidence for this includes the observation that the DvH genome encodes an abnormal number of histidine kinases. Our goal is to determine the types of protein modifications that arise in DvH and how these modifications affect the ability of DvH to survive or adapt to its environment.

This work leverages the unique resources of the Virtual Institute for Microbial Stress and Survival: Quality controlled biomass produced at LBL (Hazen lab) is used for all proteomic LC/MS/MS measurements at LBL (Keasling lab). Our initial survey of PTMs in DvH was obtained by mining these numerous proteomic LC/MS/MS data sets acquired over the course of ESPP for evidence of modified peptides. Data mining for PTMs is performed at Sandia National Labs. The searched-for modifications were determined based on literature precedence and a genome search for the existence of relevant transferases. To date we have found preliminary evidence for cysteine oxidation, lysine acetylation, and methylation of lysine and arginine. Data mining for additional PTMs is ongoing. Future work will focus on validation of these findings and determining which, if any, of these modifications play a regulatory role in DvH. Validation will require selective isolation of the proteins of interest for further characterization. Here, protein isolation is made possible through the work being performed at LBL and the University of Missouri to generate DvH mutants containing tagged versions of DvH proteins.

65 ^{GT}L

The Ech Hydrogenase is Important for Growth of *D. vulgaris* with Hydrogen

S.M. Stolyar,^{1,3*} J. Wall,^{2,3} and D.A. Stahl^{1,3}

¹University of Washington, Seattle, Washington; ²University of Missouri, Columbia, Missouri; and

³Virtual Institute for Microbial Stress and Survival, <http://vimss.lbl.gov>

Project Goals: Environmental Stress Pathway Project (ESPP) is developing computational models that describe and predict the behavior of gene regulatory networks in microbes in response to the environmental conditions found in DOE waste sites. The research takes place within the Virtual Institute for Microbial Stress and Survival (VIMSS). Based at Lawrence Berkeley National Laboratory (LBNL), VIMSS supports an integrated and multi-institutional program to understand the ability of bacteria and other microorganisms to respond to and survive external stresses. VIMSS was established in 2002 with funding from the U.S. Department of Energy Genomics:GTL Program for Rapid Deduction of Stress Response Pathways in Metal/Radionuclide Reducing Bacteria. LBNL is operated by the University of California for the U.S. Department of Energy.

One objective of the Virtual Institute for Microbial Stress and Survival (VIMSS) and the Environmental Stress Pathway Project (ESPP) is to determine the genetic and physiological basis for cooperative and competitive interactions among environmental microbial populations of relevance to the DOE. The ESPP Applied Environmental Core (AEC) and Functional Genomics Core (FGC) have identified a number of genes that may participate in cooperative interactions between sulfate reducers and methanogens under low sulfate conditions. Specifically, the gram-negative Deltaproteobacterium *D. vulgaris* is able to grow in the absence of an electron acceptor via syntrophic growth with hydrogenotrophic organisms. Despite decades of research, energy conservation in *D. vulgaris* is not well understood. The presence of multiple hydrogenases, including those located in the periplasm in all studied *Desulfovibrio* strains—and the observation that hydrogen is produced and then consumed during growth of *D. vulgaris* Miyazaki with lactate and sulfate (Tsuji&Yagi, 1980)—lead to the formulation of the hydrogen cycling hypothesis as a mechanism for energy

conservation (Odom & Peck, 1981). The availability of a completed genome sequence of *D. vulgaris* Hildenborough has since revealed genes for at least six different hydrogenases: four periplasmic and two cytoplasmic. Although several have been partially characterized biochemically and genetically, their roles in *D. vulgaris* under different growth conditions is not well understood. We examined the growth and metabolite production of an *echA* (DVU0434) *D. vulgaris* Hildenborough mutant under three different growth conditions: i) in medium amended with lactate and sulfate and ii) in medium amended with acetate, hydrogen and sulfate, and iii) in coculture the hydrogenotrophic methanogen *M. maripaludis*, lacking an electron acceptor. On lactate, the mutant demonstrated a comparable growth rate and yield to the wild type strain, but evolved more hydrogen as measured by its accumulation in the headspace during growth in batch culture. In a medium containing 5 mM acetate and an atmosphere of H₂/CO₂ (80:20), growth of the mutant was severely impaired relative to the wild type. A coculture consisting of the mutant strain and a hydrogenotrophic methanogen (*M. maripaludis*) demonstrated only slightly reduced growth rate and biomass relative to the wild type. Although this suggested some role in energy conservation, the more obvious phenotype was its greatly limited growth in monoculture with acetate, hydrogen and sulfate. Thus, the available data suggest that the primary role of the Ech Hydrogenase is oxidation of hydrogen during sulfate respiration, possibly also contributing to the production of reduced ferredoxin required for conversion of Acetyl CoA to pyruvate by pyruvate oxidoreductase, as was previously demonstrated for the homologous hydrogenases in *M. barkeri* and *M. maripaludis* (Meuer et al., 2002; Porat et al., 2006).

66 [—]GTL

Monitoring of Microbial Reduction and Reoxidation Activities in the FRC Sites using a Comprehensive Functional Gene Array

Zhili He,^{1,2,6*} Joy D. Van Nostrand,^{1,6} Liyou Wu,^{1,2} Terry J. Gentry,^{2,3} Ye Deng,¹ Christopher W. Schadt,^{2,6} Weimin Wu,⁴ Jost Liebich,² Song C. Chong,² Baohua Gu,² Phil Jardine,² Craig Criddle,⁴ David Watson,² Terry C. Hazen,^{5,6} and Jizhong Zhou^{1,2,6}

¹Institute for Environmental Genomics and Department of Botany and Microbiology, University of Oklahoma, Norman, Oklahoma; ²Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee; ³Department of Soil and Crop Sciences, Texas A&M University, College Station, Texas; ⁴Department of Civil and Environmental Engineering, Stanford University, Stanford, California; ⁵Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California; ⁶Virtual Institute for Microbial Stress and Survival, <http://vimss.lbl.gov>

Project Goals: Environmental Stress Pathway Project (ESPP) is developing computational models that describe and predict the behavior of gene regulatory networks in microbes in response to the environmental conditions found in DOE waste sites. The research takes place within the Virtual Institute for Microbial Stress and Survival (VIMSS). Based at Lawrence Berkeley National Laboratory (LBNL), VIMSS supports an integrated and multi-institutional program to understand the ability of bacteria and other microorganisms to respond to and survive external stresses. VIMSS was established in 2002 with funding from the U.S. Department of Energy Genomics:GTL Program for Rapid Deduction of Stress Response Pathways in Metal/Radionuclide Reducing Bacteria. LBNL is operated by the University of California for the U.S. Department of Energy.

A novel comprehensive functional gene microarray, termed GeoChip, has been developed. This array contains 24,243 oligonucleotide (50mer) probes and covers > 10,000 genes in >150 functional groups involved in nitrogen, carbon, sulfur and phosphorus cycling, metal reduction and resistance, and organic contaminant degradation. This array contains 24,243 oligonucleotide (50mer) probes and covers > 10,000 genes in >150 functional groups involved in nitrogen, carbon, sulfur and phosphorus cycling, metal reduction and resistance, and organic contaminant degradation. Due to the nature of functional gene sequences (highly homologous and incomplete), it is extremely challenging to select specific oligonucleotides for some functional genes using routine probe design strategies. To tackle those problems, we used the following strategies: (1) Retrieved sequences were aligned, and only the shared regions of the functional genes were used for probe design; (2) Experimentally established oligonucleotide design criteria and a novel software tool, CommOligo that was specifically developed to deal with highly similar sequences, were used for GeoChip; (3) To detect both divergent and closely related sequences, gene- and group-specific probes were designed; and (4) To increase the confidence of detection, multiple probes for each sequence or each group of sequences were designed.

The developed GeoChip is a powerful generic tool, and can be used: (1) to survey any environmental samples, such as soil, groundwater, sediments, oil fields, deep sea, animal guts, etc; (2) to study biogeochemical processes and functional activities of microbial communities important to human health, agriculture, energy, global climate change, ecosystem management, and environmental cleanup and restoration; (3) to explore direct linkages of microbial genes/populations to ecosystem processes and functions; and (4) to detect functional genes and/or organisms in a particular environment. Here, we present three related studies on the dynamics and stability of microbial genes and associated communities during bioremediation and reoxidation periods at the Oak Ridge Field Research Center (FRC) and Hanford site using the developed array.

First, Geochip was used to track the dynamics of metal-reducing bacteria and associated communities for an *in situ* bioremediation project at the FRC site in Oak Ridge. Samples were taken from different wells after ethanol injections (after day 166). During the uranium reduction period, both FeRB and SRB populations reached their highest levels at Day 212, followed by a gradual decrease over 500 days. Consequently, the uranium in groundwater and sediments was reduced, and the uranium concentrations in the groundwater were significantly correlated with the total abundance of *c*-type cytochrome genes from *Geobacter*-type FeRB and *Desulfovibrio*-type SRB. Mantel test also indicated that there was significant correlation between the differences of uranium concentrations and those of total *c*-cytochrome gene abundance or *dsrAB* gene abundance. These results suggested that *Geobacter*-type FeRB and SRB played significant roles in reducing uranium to a level below the drinking standard (<30 µg/L).

Second, the developed array was applied to study the processes of reoxidation (a period after microbial reduction) in microbial communities. Samples were taken before, during, and after a reoxidation period, during which air-saturated tap water (9-12 mg L⁻¹ DO) was injected into the FBR for a period of 77 d. DO levels in well FW101-2 and FW102-2 increased to 2 and 0.4-0.5 mg L⁻¹, respectively, during the reoxidation period and changes in the relative abundance of functional groups were apparent in both wells. For example, at 40 d post reoxidation, well FW101-2 showed an increase in the relative abundance of genes involved in ammonification, nitrification, and denitrification and a decrease for those associated with cytochromes, methane generation, N fixation, and sulfate reduction. During the post-reoxidation (Day 77), the relative abundance of ammonification, denitrification, and ammonia oxidation genes had returned to pre-oxidation levels in FW101-2, while genes for methane oxidation and nitrate reduction increased and those for nitrite reduction decreased.

In addition, the developed GeoChip was used to evaluate functional communities at a lactate-fed chromium reduction system at the Hanford site. Samples were taken from different depths within injection and extraction wells. Extraction well samples showed higher numbers of functional genes than the injection well at the same depth. Within the extraction well, abundance decreased with depth. However, the relative abundance of chromium resistance gene increased with depth in this same well.

The developed GeoChip is the most comprehensive functional gene array for environmental studies so far, but due to exponential increases in the numbers of genes and the number of sequences for each gene, we expect to continuously update the array to reflect the sequence information currently available in public databases and personal collections if possible. Thus, we are working on the third generation array, which covers almost three times more gene sequences, and has more features. For example, *gyrB* has been added for phylogenetic analysis. In addition, a software package (including databases) has been developed for sequence retrieval, probe and array design, probe verification, array construction, array data analysis, information storage, and automatic update.

67 ^{GT}L

Towards High-Throughput and High Sensitivity Approaches for Uncovering Total Environmental Gene Expression Patterns

Zamin Yang,^{1,3} Christopher W. Schadt,^{1,3*} Terry Hazen,^{2,3} and Martin Keller^{1,3}

¹Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee; ²Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California; and ³Virtual Institute for Microbial Stress and Survival, <http://vimss.lbl.gov/>

Project Goals: Environmental Stress Pathway Project (ESPP) is developing computational models that describe and predict the behavior of gene regulatory networks in microbes in response to the environmental conditions found in DOE waste sites. The research takes place within the Virtual Institute for Microbial Stress and Survival (VIMSS). Based at Lawrence Berkeley National Laboratory (LBNL), VIMSS supports an integrated and multi-institutional program to understand the ability of bacteria and other microorganisms to respond to and survive external stresses. VIMSS was established in 2002 with funding from the U.S. Department of Energy Genomics:GTL Program for Rapid Deduction of Stress Response Pathways in Metal/Radionuclide Reducing Bacteria. LBNL is operated by the University of California for the U.S. Department of Energy.

In recent years, tremendous progress has been made in understanding microbial communities due to emergence of newly developed genomics-based technologies. Current technologies that have been applied to environmental samples for RNA transcriptional profiling include RT-PCR and functional gene microarrays using total RNAs. While these methods have provided considerable insights, they bear significant limitations that prevent their application in a high throughput manner to *de novo* communities. Both methods require background genomic information to allow for design of specific primers and/or microarray probes. Consequently, these methods can only reveal transcription activity of targeted conserved genes that are first surveyed by PCR and sequencing methods, or those that are obtained through comprehensive metagenomic shotgun sequencing. Thus each of these methods

have high upfront costs in time, effort and materials. To circumvent this limitation, we are developing a method involving direct sequencing cDNA from the environment samples utilizing a high throughput sequence analysis system such as Bio454. These types of tools will be especially useful in understanding the basis for microbial survival under extreme environmental stressors which is a primary goal of the VIMSS:ESPP project.

Since about 80% of total RNA from microbial organisms consists of ribosomal rRNAs it is crucial to first remove rRNAs as completely as possible without degrading mRNA quality and quantity prior to HT sequence based screening. Since bacteria primarily produce mRNAs without a poly-A tail and thus cannot be enriched using oligo-dT methods, as first step to the application of HT sequencing we have compared three different methods to remove rRNAs and enrich mRNAs of *Desulfovibrio vulgaris* Hildenborough samples. The first method utilizes biotin modified oligos complementary to conserved regions in 16S & 23S rRNA and subtractive hybridization with streptavidin-coated magnetic beads. The second uses a commercially available exonuclease that specifically digests rRNAs bearing a 5' monophosphate group. The third method uses two rounds of reverse transcription, where rRNAs are first reverse transcribed with multiple universal primers for 16S & 23S RNAs, subsequently the RNA/DNA hybrids and cDNA are removed by sequential digestion with RNaseH and DNaseI, and the enriched mRNAs are then reverse transcribed using random primers.

We evaluated these three methods by comparing disappearance of the 16S and 23S bands via electrophoresis, and their effect on mRNA quality by analysis of transcription levels of control (total RNA) vs. enriched mRNA as measured by a whole genome microarray. While all three methods were able to significantly enrich mRNA from rRNA, the microarray analysis revealed differences in measured mRNA levels. In control vs. control (unenriched) hybridizations, less than 0.2 % of genome (5 of 3604 total genes) exhibited significant ($P < 0.05$) changes in the levels of their transcripts. Enriched mRNAs from the first two methods generated on average more genes with altered transcript levels compared to untreated total RNA, with 19 genes (0.5%) for the exonuclease method & 74 genes (2%) for subtractive hybridization exhibiting significantly different than controls. In fact, each of these methods appeared to increase the sensitivity of detection, as average signal intensity's corrected for background were 16% higher for the exonuclease method and 113% for the subtractive hybridization method. Microarray comparisons for the third method are currently under analysis. After completing this initial evaluation, we will use each of these methods to construct cDNA libraries for HT sequencing with the 454 to further optimize and validate this approach in single species, as well as make comparisons of HT sequence based methods with existing microarrays. Subsequent validation and application of the developed methods will be performed on mixed cultures (*Desulfovibrio* & *Methanococcus*) and incorporation of amplification steps. The developed tools will then be deployed to understand microbial survival in stressed environmental systems.

Experimental and Computational Approaches to Enhance Proteomics Measurements of Natural Microbial Communities

Nathan VerBerkmoes,^{1*} Mark Lefsrud,¹ Chongle Pan,¹ Brian Erickson,¹ Manesh Shah,¹ Chris Jeans,² Steven Singer,² Michael P. Thelen,² Vincent Deneff,³ I. Lo,³ **Jillian Banfield**³ (jill@eps.berkeley.edu), and Robert L. Hettich¹

¹Oak Ridge National Laboratory, Oak Ridge, Tennessee; ²Lawrence Livermore National Laboratory, Livermore, California; and ³University of California, Berkeley, California

Project Goals: The interdisciplinary research in our GTL project is motivated by the desire to understand how microbial communities assemble, the interplay amongst community members, and the mechanisms of microbial adaptation and evolution. Chemoautotrophic acid mine drainage (AMD) microbial biofilms have proven particularly tractable for these analyses due to their comparatively low species richness. The goal of this subproject is to develop experimental and computational approaches for the comprehensive characterization of the proteome of the AMD system to investigate the nature of the gene expression and conservation amongst the various microbial members of this consortium. Proteomic information will be integrated with genomic and biochemical datasets to help elucidate the structure and activity of microbial communities in their natural environmental context.

Microorganisms comprise the majority of extant life forms and play key roles in a wide variety of health and environmental processes, yet little is known about the nature and driving forces of their diversification. Most research to date has focused on physiological and genomic characterization of a relatively small number of isolated microbial species maintained in monoculture. While providing crucial insights to connect genes and function, these studies are unable to capture some aspects of the organism's behavior in its natural environment. Perhaps the greatest knowledge gap is in the understanding of how microorganisms function within natural multi-species consortia. Recently, genomic characterization has been extended to natural microbial communities, opening the way for cultivation-independent analysis of microbial activity in environmental context.

We have used proteomic methods to analyze biofilm samples from acid mine drainage (AMD) biofilms previously characterized by cultivation-independent genomic methods (Ram, Science, 2005). This community consists of bacteria, archaea, and eukaryotes that have adapted to survive in an extreme environment (pH <1.0, high metal content, high temperature). LC/LC-MS/MS based "shotgun" proteomics with LTQ and LTQ-Orbitrap technologies identified over 2,000 proteins, yielding functional information about each of the five dominant species. The recent acquisition of additional genome data indicated the presence of two major strain variants of the dominant *Leptospirillum* II species. Detailed proteomic measurements provided the first evidence for large-scale genome recombination. The key methodological advance here is the finding that it is possible to deduce the sequences of gene variants, so long as genomic data from relatively closely related organisms are available.

Based on this initial work, we have established an experimental/computational proteomics pipeline at ORNL for the microbial community samples. To date, we have characterized six distinct biofilm samples from different regions in the Richland mine. In each sample, we were able to measure between 2,000 – 3,000 non-redundant proteins, including 1,000 – 1,200 unknown proteins (i.e. proteins that are predicted from the community genomic sequence whose existence has not previ-

ously been confirmed and which have no functional annotations). In many cases, these unknown proteins were identified in multiple samples, verifying the notion that they play key functional roles in the community operation, while other unknown proteins were found in individual biofilms suggesting potential spatial and/or temporal expression. We are in the process of examining the correlation between the fluorescent *in-situ* hybridization (FISH) data (which provides information about the microbial species abundance in each sample) and the proteome data (which provides protein information primarily from the most abundant organisms in each sample). In general, the proteomic information is providing important feedback for the genomic annotation and assembly. Whereas the genome information reveals genetic potential for the community members, the proteome data indicates the nature of the gene expression and conservation in the various species.

We have been able to acquire and integrate a high performance mass spectrometer into our proteomics pipeline this last year. In particular, this hybrid linear trapping quadrupole-Orbitrap (LTQ-Orbitrap) MS provides high resolution, accurate mass measurements on both parent and fragment ions in a high-throughput data-dependent fashion and on liquid chromatography time scales. The improved dynamic range of proteome measurement achievable with this instrumentation, when combined with the high resolution metrics stated above, provide enhanced capabilities for deeper and more accurate proteome measurements of these very complex samples. The LTQ-Orbitrap can routinely give mass accuracies on parent peptides with millimass accuracies (<5ppm mass error) while still permitting rapid MS/MS scans in the LTQ. With a 24-hour measurement, it is routine to measure 1500-2500 proteins from AMD samples. The ultra high mass accuracy combined with LTQ MS/MS spectra virtually eliminates false positives. A somewhat slower mode of operation is to collect both full-scan and MS/MS scans in the Orbitrap. This allows for very high mass accuracy on fragment ions, which facilitates *de-novo* sequencing. Even in this lower duty cycle mode of operation, 1000-1500 proteins can be identified in 24-hours and high abundance, high quality MS/MS spectra can be extracted and analyzed by *de-novo* sequencing algorithms.

We have also upgraded our existing Fourier transform ion cyclotron resonance mass spectrometer (FTICRMS) to incorporate a new electrospray ion source to enable advanced protein/peptide fragmentation techniques (infrared multiphoton dissociation (IRMPD) and electron capture dissociation (ECD)). These methods are complementary to the more conventional collisional fragmentation and are especially useful for larger peptides and proteins. In collaboration with our LLNL colleagues, we have been able to use this MS to investigate intact cytochrome proteins isolated from the extracellular biofilm fraction. Not only were we able to investigate the molecular mass heterogeneity of the various forms of cytochrome 579, but we also were able to use the IRMPD technique along with *de novo* sequencing approaches to discover and validate strain variations in the amino acid sequences. This provides important information that can be used to assist the “bottom-up” peptide identifications of strain variation in this natural community. We have begun work to use the new LTQ-Orbitrap technique for proteome quantification, achieved by isotopically labeling biofilm samples grown in laboratory bioreactors at UC-Berkeley. By comparing non-labeled and labeled biofilms, it should be possible to collect more definitive information about the relative abundance changes of proteins as a function of different biofilm growth states.

The explosion of experimental biofilm data this last year has prompted increasing bioinformatic needs. We have optimized our computational approach for data mining, interpretation, and dissemination. See http://compbio.ornl.gov/biofilm_amd/ and http://compbio.ornl.gov/biofilm_amd_recombination/ as examples. In addition, a new “peptide viewer” has been constructed to permit a more detailed examination of peptide sequence variation across different biofilm samples. A detailed informatic study was conducted to ascertain the level of strain variation that could be deciphered by MS-based proteomic measurements. The key concern motivating this study was the fact that even

a single amino acid variation is sufficient to prevent *peptide* identifications; thus, it was necessary to evaluate what level of sequence variation at the amino acid level would preclude *protein* identifications. Further research has focused on false positives at different filtering levels for the AMD system and how this can be improved with high mass accuracy. The continuing goal is to limit both false positives and false negatives while obtaining as complete proteome coverage as possible.

On-going research continues to be directed at pushing the experimental and computational capabilities for deep and accurate proteome characterization in complex microbial communities. The limited complexity of the acid mine drainage system is the perfect system to begin to develop and evaluate these tools. In particular, we have learned a great deal about how to measure and decipher strain variation in microbial consortia. This work provides important information about how to deal with strain variation in complex systems of interest for both human health and environmental applications.

This research sponsored by the U.S. DOE-BER, Genomics:GTL Program. Oak Ridge National Laboratory is managed by UT-Battelle, LLC, for the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

69 ^{GTL}

Strain-Resolved Proteogenomics-Enabled Ecological Study of Natural Microbial Communities Associated with Acid Mine Drainage Formation

V.J. Denef^{1*} (vdenef@berkeley.edu), N.C. VerBerkmoes,² P. Wilmes,¹ M. Shah,² D. Goltsman,¹ I. Lo,¹ G. DiBartolo,¹ L. Kalnejais,¹ B.J. Baker,¹ G.W. Tyson,¹ J.M. Eppley,¹ E.A. Allen,¹ R.L. Hettich,² M.P. Thelen,³ and J.F. Banfield¹

¹University of California, Berkeley, California; ²Chemical Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee; and ³Biosciences Directorate, Lawrence Livermore National Laboratory, Livermore, California

Project Goals: The interdisciplinary research in our GTL project is motivated by the desire to understand how microbial communities assemble, the interplay amongst community members, and the mechanisms of microbial adaptation and evolution. Chemoautotrophic acid mine drainage (AMD) microbial biofilms have proven particularly tractable for these analyses due to their comparatively low species richness. Our goal is to use this system to develop cultivation-independent community genomic and proteomic methods and apply them to study the structure and activity of microbial communities in their natural environmental context.

Microbially promoted dissolution of metal sulfides leads to the formation of acid mine drainage (AMD), a major environmental problem associated with energy resources. It is also a process that underpins bioleaching-based metal recovery and coal desulfurization and mercury removal. Due to a limited spectrum of energy harvesting and metabolic opportunities, the chemoautotrophic microbial communities that populate acid mine drainage systems tend to have low species richness and thus are particularly amenable to high-resolution ecological analyses. Our approach is to use genomic data from two spatially and temporally separated natural microbial communities sampled at the Richmond mine (Iron Mountain, Redding, CA) to characterize the proteomes of multiple biofilms and to correlate the activities of organisms with community structure and environmental conditions.

To date, extensive community genomic data have been obtained from an air-solution interface biofilm (pH 0.83, 42 °C; Tyson et al., 2004) at the 5-way location and a subaerial biofilm (pH 1.1, 41 °C; Lo et al., resubmitted) at the UBA location. The *Leptospirillum* group II species from the 5-way CG and UBA datasets differ by 0.3% at the 16S rRNA gene level. The genomes are highly syntenous and share 83% and 76% orthologs (measured relative to the UBA 5-way CG gene inventories, respectively) with 95.24% average amino acid sequence identity (median = 96.69 %).

In order to evaluate the physiological and ecological significance of the *Leptospirillum* group II variants, we designed strain-specific fluorescent *in situ* hybridization (FISH) probes targeting the 23S rRNA gene to enable us to correlate the *Leptospirillum* group II type with environmental conditions. We detect no connection between genome type and pH, temperature, or ionic strength. However, the UBA type dominates in the early and middle stages of biofilm development whereas the 5-way CG type predominates in late successional stages when Archaea and fungi become important. Characterization of intact biofilm cross sections suggests that *Leptospirillum* group II most intensively colonize parts of the biofilm in direct contact with Fe-rich solutions whereas Archaea partition in the upper biofilm regions and show some association with fungal filaments. In contrast, *Leptospirillum* group III, the only nitrogen fixer yet identified in the system, is distributed throughout biofilms as single cells or microcolonies.

The reconstructed *Leptospirillum* group III genome encodes ~3,000 genes, with an average amino acid identity to *Leptospirillum* group II UBA type of ~58% (~1,800 orthologs). Compared to *Leptospirillum* group II, most biofilms contain a single near-clonal *Leptospirillum* group III type. However, genomic analyses reveal the same variants of the majority of *Leptospirillum* group III genes in all samples analyzed. Heterogeneity in *Leptospirillum* group II and III is typically in the form of differences in gene content, especially in putative prophage or plasmid-like regions. Analysis of genomic datasets revealed that recombination is an important process shaping archaeal populations in AMD biofilms. Quantitative genome-wide analyses indicate the frequency of recombination has a log-linear dependence on sequence divergence, with a significant discontinuity between rates within vs. between genomic clusters. The role of recombination in shaping the genetic potential of *Leptospirillum* group II was uncertain. Comparisons of the *Leptospirillum* group II UBA vs. 5-way CG genomes revealed regions of 10s to 100s kb in length and comprising a total of 421 genes sharing essentially identical nucleotide sequence. Excluding the subset attributed to integration of identical phage or IS elements, evidence suggests that the UBA and 5-way CG *Leptospirillum* species were shaped by recent homologous recombination between two organism types, followed by selective sweeps. Recombination is typically documented by sequence comparisons involving a few genes from organisms obtained in pure culture. In this study, we used shotgun proteomics to map, genome-wide, strain-specific expressed protein variants in a third community for which no genome sequence is available. Results reveal a *Leptospirillum* group II population dominated by a single organism with a genome of predominantly UBA type genes, but with chromosomal regions tens to hundreds of kilobases in length that derived from the 5-way CG genome type (confirmed by multi-locus sequence typing of isolates and uncultivated natural consortia). For both bacteria and archaea, within-gene recombination is an important source of new gene variants. Results suggest that formation of hybrid genome types is important in fine-scale environmental adaptation.

The ability to distinguish between proteins that differ by as few as a single amino acid enables strain-specific proteomics studies to resolve the behavior of closely related members of natural communities. Proteomic analyses of two mature biofilms that are colonized by both *Leptospirillum* types are underway. Replicate samples of biofilms representing different growth stages, as well as a biofilm showing evidence suggestive of phage predation, are in line for analysis in the next few months.

Funding was provided by the DOE Genomics:GTL Program under grant number DE-FG02-05ER64134 (Office of Science), the NSF Biocomplexity Program, and the NASA Astrobiology Institute.

70 ^{GTL}

A Novel Iron Oxidase Isolated from an Extremophilic Microbial Community

Steven W. Singer,^{1*} Christopher Jeans,¹ Jason Raymond,¹ Adam Zemla,^{1,2} Nathan C. VerBerkmoes,³ Robert L. Hettich,³ Clara Chan,⁴ **Jill Banfield**,⁴ and Michael P. Thelen¹ (mthelen@llnl.gov)

¹Chemistry, Materials and Life Sciences and ²Computation Directorates, Lawrence Livermore National Laboratory, Livermore, California; ³Chemical Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee; and ⁴Department of Earth and Planetary Sciences, University of California, Berkeley, California

Project Goals: The interdisciplinary research in our GTL project is motivated by the desire to understand how microbial communities assemble, the interplay amongst community members, and the mechanisms of microbial adaptation and evolution. Chemoautotrophic acid mine drainage (AMD) microbial biofilms have proven particularly tractable for these analyses due to their comparatively low species richness. Our goal is to use this system to develop cultivation-independent community genomic and proteomic methods and apply them to study the structure and activity of microbial communities in their natural environmental context.

Proteins isolated directly from uncultivated microbial populations represent critical functional components of community structure and metabolism. Our investigation of a chemoautotrophic microbial community, typified by biofilm formation, iron oxidation and environmental acidification, has resulted in the identification of a large fraction of abundant proteins that do not correlate with any recognized sequences. As there are no systematic methods at hand to analyze “hypothetical proteins”, we have developed an approach towards functional determination using proteogenomic, structural modeling and biochemical tools, with an initial focus on novel iron oxidizing proteins.

Two very abundant novel proteins isolated from acidophilic biofilms collected in Iron Mountain’s Richmond Mine (described by JF Banfield in other presentations) were identified as *Leptospirillum* group II gene products and characterized as cytochromes with unique and unusual properties. A sulfuric acid wash of biofilm samples liberates a major 16 kDa cytochrome with an unusual α -band absorption at 579 nm, Cyt₅₇₉ [Ram et al., *Science* 2005]. Using Cyt₅₇₉-specific antibodies, electron microscopic imaging of biofilm thin sections localizes the cytochrome to the outer cell surface of *Leptospirillum* bacteria. Purification from different microbial community samples unexpectedly indicated several different variations in protein mass and isoelectric points of Cyt₅₇₉. Further genomic sequencing of field samples indicated that several paralogs and variants of Cyt₅₇₉ are present in distinct *Leptospirillum* strains, and MS proteomics confirmed that at least five of these variants occur in different amounts depending on the location and growth stage of the biofilm examined. Although the variations are due to one or more amino acid substitutions, multiple truncations at both ends of Cyt₅₇₉ proteins were also determined, indicating processing that is perhaps important for cytochrome placement and function. Solution measurements of the purified protein point to a single domain, monomeric, α -helical structure. This is important in its interactions with other cytochromes involved in the oxidation of Fe(II) and electron transfer.

In addition to the small soluble Cyt₅₇₉, membrane fractions of biofilm samples are dominated by a 60 kDa, heme-bearing protein with an α -band absorption at 572 nm. This cytochrome, Cyt₅₇₂, is located in the outer membrane of *Leptospirillum* group II, and again there are several variants and possible paralogs present in the community genome data. Cyt₅₇₂ genes fall within a recombination hotspot between two strains of *Leptospirillum* group II, perhaps indicating the selection of variants of a protein essential to survival by its close association with changing geochemical conditions. Solution measurements of this purified cytochrome indicate a two-domain, oligomeric, and predominantly β -stranded structure.

Although both cytochromes 579 and 572 are so far unique to the *Leptospirilla* bacteria, a new “domain fusion” approach to structural modeling was used to test their similarity with known protein structures. Surprisingly, both cytochromes appear to have structural homology with bacterial nitrite reductases such as NirB. These are outer membrane proteins with homodimeric, two-domain structures. A small N-terminal heme binding domain is a good template for the entire Cyt₅₇₉ polypeptide and also a relatively good fit to the N-terminal domain of Cyt₅₇₂. The large C-terminal domain of these nitrite reductases are 8-bladed β -propeller structures observed in many other proteins including those involved in signal transduction; this domain provides a structural scaffold for modeling the C-terminal domain of Cyt₅₇₂. These predictions are supported by solution measurements of purified cytochromes 579 and 572, including circular dichroism, protease digestion, size exclusion chromatography, and cytochrome oxidation by nitrite.

A general mechanistic model involving iron oxidation has resulted from our studies. Fe(II) oxidation by isolated Cyt₅₇₂ occurs readily at pH 0.95 – 3, whereas Cyt₅₇₉ is less reactive at low pH. Under specified conditions, Cyt₅₇₂ transferred electrons to Cyt₅₇₉, perhaps representing an initial step in energy flow from the environment to the biofilm. Interestingly, a recently sequenced acidophilic bacterium contains a distinct operon of cytochromes that are homologous to cytochromes 579, 572, and a cytochrome oxidase. This indicates a coordinated regulation of these novel cytochrome genes and links these proteins to the generation of toxic mine drainage by acidophilic biofilm communities.

71 ^{GT}L

Functional Analysis of Protein Phosphorylation in *Shewanella oneidensis* MR-1

C. Giometti^{1*} (csgiometti@anl.gov), G. Babnigg,¹ A. Beliaev,² G. Pinchuk,² M. Romine,² and J. Fredrickson²

¹Argonne National Laboratory, Argonne, Illinois and ²Pacific Northwest National Laboratory, Richland, Washington

Project Goals: This project is a component of the *Shewanella* Federation and as such contributes to the overall goal of applying the tools of genomics, leveraging the availability of genome sequence for 18 additional strains of *Shewanella*, to better understand the ecophysiology and speciation of respiratory-versatile members of this important genus. To understand these systems the SF is using genome-based approaches to investigate *Shewanella* as a system of integrated networks; first describing key cellular subsystems—those involved in signal transduction, regulation, and metabolism—then building towards understanding the function of whole cells and, eventually, cells within populations. As a general approach, the SF is collectively employing complementary “top-down” bioinformatics-based genome functional predictions, high-throughput expres-

sion analyses, and functional genomics approaches to uncover key genes as well as metabolic and regulatory networks. The “bottom-up” component employs more traditional approaches including genetics, physiology and biochemistry to test or verify predictions. This information will ultimately be linked to analyses of signal transduction and transcriptional regulatory systems and used to develop a linked model that will contribute to understanding the ecophysiology of *Shewanella* in redox stratified environments.

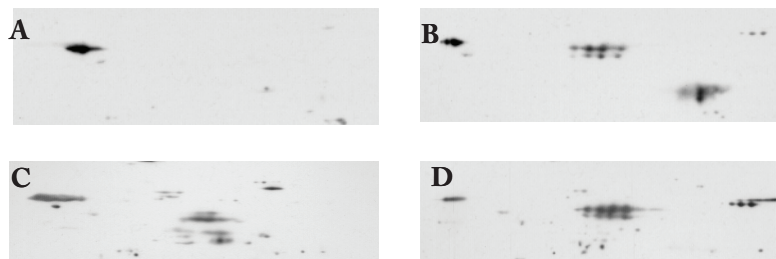


Figure 1. Western blot analysis of *S. oneidensis* proteins revealed proteins phosphorylated on serine or threonine residues. A, phosphothreonine proteins from MR-1 grown aerobically; B, phosphothreonine proteins from MR-1 grown anaerobically with fumarate; C, phosphoserine proteins from MR-1 grown aerobically; D, phosphoserine proteins from MR-1 grown anaerobically with fumarate.

Protein phosphorylation plays an important role in the regulation of cell physiology in both prokaryotes and eukaryotes. Using a suite of proteomics tools, including affinity chromatography, gel electrophoresis, Western blotting, and tryptic peptide mass analysis, we have identified a set of phosphoserine and phosphothreonine proteins expressed by *Shewanella oneidensis* MR-1 cells. Differential expression of a subset of these phosphoproteins, including pyruvate formate lyase and fumarate reductase (Figure 1) has been observed in cells grown with different electron acceptors. These results suggest that the phosphorylation status of these proteins is involved in regulating carbon and energy metabolism in *S. oneidensis* MR-1. To determine the functional significance of the observed differential phosphorylation in response to electron acceptor availability, protein phosphorylation in MR-1 strains deficient in key respiratory proteins, the tetraheme cytochrome c, CymA, and fumarate reductase, FccA, is being investigated. These mutants, generated as part of the *Shewanella* Federation efforts to characterize MR-1 carbon and energy metabolism, are being grown in parallel with MR-1 wild type with different electron acceptors. The phosphoserine and phosphothreonine protein profiles of these cultures are then generated by using Western blot analysis, and quantitative comparative analysis is used to detect significant differences in the phosphoprotein profiles. Characterization of the phosphorylation sites is being done using mass spectrometry. In addition, studies are being done to further characterize the differential phosphorylation of pyruvate formate lyase observed in response to oxygen availability and the phosphorylation of fumarate reductase, known to be a periplasmic protein in *S. oneidensis*.

This work is funded by the U.S. DOE Office of Biological and Environmental Research under Contract No. DE-AC02-06CH11357.

72 ^{GTL}

Enriching Metabolic Function Predictions for *Shewanella oneidensis* MR-1 with Growth and Expression Studies

Margrethe H. Serres^{1*} (mserres@mbi.edu), Margaret F. Romine,² Mike E. Driscoll,³ Tim S. Gardner,³ Natalia Maltsev,⁴ Miriam Land,⁵ Andrei Osterman,⁶ Mary Lipton,² and LeeAnn McCue²

¹Marine Biological Laboratory, Woods Hole, Massachusetts; ²Pacific Northwest National Laboratory, Richland, Washington; ³Massachusetts Institute of Technology, Boston, Massachusetts; ⁴Argonne National Laboratory, Argonne, Illinois; ⁵Oak Ridge National Laboratory, Oak Ridge, Tennessee; and ⁶Burnham Institute, La Jolla, California

Project Goals: This project is a component of the *Shewanella* Federation and as such contributes to the overall goal of applying the tools of genomics, leveraging the availability of genome sequence for 18 additional strains of *Shewanella*, to better understand the ecophysiology and speciation of respiratory-versatile members of this important genus. To understand these systems the SF is using genome-based approaches to investigate *Shewanella* as a system of integrated networks; first describing key cellular subsystems—those involved in signal transduction, regulation, and metabolism—then building towards understanding the function of whole cells and, eventually, cells within populations. As a general approach, the SF is collectively employing complimentary “top-down” -bioinformatics-based genome functional predictions, high-throughput expression analyses, and functional genomics approaches to uncover key genes as well as metabolic and regulatory networks. The “bottom-up” component employs more traditional approaches including genetics, physiology and biochemistry to test or verify predictions. This information will ultimately be linked to analyses of signal transduction and transcriptional regulatory systems and used to develop a linked model that will contribute to understanding the ecophysiology of *Shewanella* in redox stratified environments.

The genome sequence *Shewanella oneidensis* MR-1 (MR-1) was released in 2002 (1) and reannotated one year later (2). MR-1 is the model organism being used by the *Shewanella* Federation team to develop a better understanding of the ecophysiology and speciation of respiratory-versatile members of the *Shewanella* genus. These genome annotations have served as an essential resource for designing experiments to investigate MR-1 function and for interpreting experimental results generated from high through-put analyses such as microarray and global proteomics. In order to improve the accuracy of these initial annotations we have applied various new bioinformatics tools including Puma2, Gnare, and SEED as well as information from 15 new *Shewanella* genome sequences to improve functional predictions based on sequence information. In addition, we have used proteome data generated by the *Shewanella* Federation to improve ORF calls and to validate predictions of protein processing. The new ORF predictions were used to design an MR-1 specific Affymatrix chip and to create a modified protein sequence file for proteome analysis. They have also been used to improved predictions of operon structure and cellular localization of proteins.

Predicted functions have been used to generate a picture of the intermediate metabolism of MR-1 (3). Analogous to the pathways described in the BioCyc database as well as the literature, we have made approximately 150 pathway predictions, and we are constructing a *Shewanella* pathway database (SheonDB) for MR-1 using the Pathway Tools software (4). In addition public metabolic databases i.e. the KEGG map are being consulted.

We are now seeking to further improve the accuracy of our current predictions by interrogating additional experimental resources, such as microarray analyses of expression patterns and growth

phenotypes. A comparative analysis of cells grown aerobically with one of 5 different carbon sources (casamino acids, N-acetyl-D-glucosamine, inosine, pyruvate and lactate) supported many previously reported pathway assignments (3) and suggested alternative assignments for others. Such a case was seen for the degradation of the branched chain amino acids of leucine, valine, or isoleucine. The first step, a deamination reaction, was proposed to be catalyzed by either the branched-chain-amino-acid aminotransferase IlvE (SO0340) or by the leucine dehydrogenase Ldh (SO2638). The microarray data suggests that SO2638 catalyzes this reaction during growth on casamino acids generating NADH and ammonia as byproducts. Also, according to the KEGG map the third step in the conversion of these amino acids to the respective methyl butanoyl-CoA intermediate (EC 2.3.1.-) was missing an assignment for MR-1. The prediction that SO2341 (EC 2.3.1.168) carried out this step (3) was supported by the microarray data. By contrast our earlier predictions for subsequent degradative steps in valine and isoleucine are not supported by the microarray data. We originally predicted that the sequential conversion of isobutyryl-CoA (valine degradation) to propionyl-CoA and acetyl-CoA was mediated via EC 1.3.99.12 (SO0021), EC 4.2.1.17 (SO0021, SO1681), EC 3.2.1.4 (SO0020), EC 1.1.1.31 (SO1682), and EC 1.2.1.27 (SO1678). Combined results from genome neighborhood analysis, literature analysis, and microarray data suggest that instead the process is mediated by EC 1.3.99.12 (SO1679), EC 4.2.1.17 (SO1681), EC 3.1.2.4 (1680), EC 1.1.1.31 (SO1682), and EC 1.2.1.27 (SO1678).

Our results suggest that integration of bioinformatics analysis with experimental data interrogation can provide improved annotations and hence can more effectively drive subsequent experimental design and data interpretation. Additional examples of how we have applied microarray data to predict new functions will be presented in the poster.

References

1. Heidelberg, J.F. et al. (2002) Genome sequence of the dissimilatory metal ion-reducing bacterium *Shewanella oneidensis*. *Nat Biotechnol.* 20:1118-23.
2. Darselia N. et al. (2003) Reannotation of *Shewanella oneidensis* genome. OMICS. 7:171-5.
3. Serres M.H., Riley, M. (2006) Genomic analysis of carbon source metabolism of *Shewanella oneidensis* MR-1: Predictions versus experiments. *J Bacteriol.* 188(13):4601-9.
4. Karp, P., Paley, S., Romero, P. (2002) The Pathway Tools Software. *Bioinformatics* 18:S225-32.

73 ^{GTL}

Proteomics Technologies Advance the Understanding of Microbial Systems Allowing for In-Depth Characterization of Microbes Important for Bioenergy Production, Bioremediation and Carbon Sequestration and Cycling

Mary S. Lipton* (mary.lipton@pnl.gov), Joshua Turse, Stephen Callister, Kim K. Hixson, Xuixia Du, Angela Norbeck, Samuel Purvine, Feng Yang, Margie F. Romine, Carrie D. Nicora, Joshua Adkins, Richard D. Smith, and Jim K. Fredrickson

Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington

Project Goals: The understanding of how cells function at the systems level will greatly benefit from the development of new approaches capable of making global measurements of protein expression (proteome). The aim of this project is to apply new capabilities that are being devel-

oped for quantitative and high throughput proteomic measurements, based primarily upon high resolution separations combined with the unique high field Fourier transform ion cyclotron resonance mass spectrometry technology developed at PNNL. The project focuses on biological applications involving studies of the proteomes of several microorganisms including *Shewanella oneidensis* MR1, *Deinococcus radiodurans* R1, *Rhodobacter sphaeroides*, *Pelagobacter ubique*, *Caulobacter crescentus* and *Geobacter* species that are of interest to the DOE Genomics:GTL program. Each organism plays a significant role in bioremediation, carbon sequestration energy production or furthering the understanding of biological systems. These efforts are proceeding in collaboration with biologists with expertise in each of these organisms, and in a close collaboration with each of the biologists, the proteomic data is translated into biological implications of changes in cellular stress or state. The focus of each of the organisms is listed below as well as the collaborator that serves as the biological lead for these subprojects.

Collaborators (Laboratory leads): Michael Daly (USUHS), Sam Kaplan (UT-Houston Medical School), Tim Donohue (University of Wisconsin-Madison), Lucy Shapiro (Stanford University), Stephen Giovannoni (Oregon State University), Derek Lovley (University of Massachusetts), Andre Osterman (Burnham Institute), Pavel Pevzner (UCSD)

Exploiting microbial function for purposes of bioremediation, energy production, carbon sequestration and other missions important to the DOE requires in-depth systems level knowledge of the molecular components of the cell that confer function. Inherent to developing a systems level understanding is comprehensive characterization of cellular proteins and how their abundance, location and modification state respond to changing conditions. Recent advances in proteomics technologies at PNNL have allowed the determination of shifts in relative abundance, localization, interactions, and post-translational modifications of cellular proteins. We will present illustrations of how the high throughput technologies at PNNL have been utilized to elucidate these post-transcriptional events in microbial cells.

Global proteomics is now capable of comprehensively identifying cellular proteins. This information has proved useful in the annotation of microbial genomes. For example, global proteomics has validated hypothetical proteins, n- and c- terminal sequences, various post-translational processing and signal peptide cleavage events. In *Shewanella oneidensis*, a blind search methodology was applied to identify peptides arising from post-translational enzymatic processing or chemical modification. Standard peptide identification methods were then used to refine protein-coding gene starts and predictions of signal peptidase cleavage greatly improving predictions of which proteins are secreted into the cell envelope. In addition to confirming many prior discoveries, this novel application of proteomics revealed 1) predicted to arise by natural mutation/variation in the cell population were in fact not the result of sequencing mistake 2) several new start codons including one instance of a rare ATA start codon, and 3) 390 instances of chemical modification of proteins.

While comprehensive identification of cellular proteins has important applications, , robust measurements of abundance are needed. The technique that shows the most promise when used in combination with the AMT tag approach is the utilization of absolute peak intensity from high-resolution FTICR instruments. Compared to stable isotope labeling and spectrum counting, this method produces quantitative data with a higher confidence from a single run, thus allowing for a statistical handling of multiple technical and biological replicates. Such analyses provide the global determination of quantitative protein response to the culture conditions or environmental condition. Changes in protein abundance levels were investigated in *Geobacter sulfurreducens* grown on fumarate or Fe(III) citrate. The abundance of proteins was determined and the results compared to identify proteins associated with these distinct modes of anaerobic respiration. Among the proteins that changed,

91 *c*-type cytochromes were identified. Relative abundance of some *c*-type cytochromes varied markedly with different growth conditions.

While the characterization of steady state cultures focus on the static survey of the cell, time course studies allow for understanding the dynamics of a system. High-throughput proteomics technologies make it possible to examine the dynamics and regulatory mechanism of biological pathways via measuring the protein expressions at multiple time points. For example, in *Rhodobacter sphaeroides*, the response of the bacteriochlorophyll production pathway increase over time as cells transition from an aerobic to a photosynthetic state suggesting the importance of this pathway in the synthesis of the photosynthetic reaction center.

While the characterization of protein abundance allows an understanding of how cells respond to environmental conditions, protein location within cells can change as a function of time and conditions. The global determination of these sub-cellular protein localizations is an ability that is unique to proteomics methodologies. We present results for *Rhodobacter sphaeroides* that characterize the proteome of aerobic and photosynthetic cell cultures by utilizing: 1) proteins extracted from whole cell lysate, soluble, insoluble, and global fractions, and 2) proteins extracted from sub-cellular fractions that include cytoplasm, cytoplasmic membrane, periplasm, outer membrane, and chromatophore. Additionally, the application of informatics techniques to those proteins that are assigned to multiple locales can aid in the determination of potential protein interaction partners since the co-localization events could arise from the formation of interactions among proteins.

Post-translational modifications are important components of protein function. Characterization of these modifications is another application that is unique to proteomics technologies. Types of modifications can vary from the addition of large moieties like hemes to addition of small ligands, such as phosphorylation. Heme modification plays an important role in electron transfer and enzyme catalysis, while phosphorylation at histidine and aspartate residues of the response regulator is essential to regulate the signal transduction pathway in the two-component system. Oxidation of proteins is often a response to a stress on the cell. The use of specialized separation or enrichment schemes in combination with high-resolution mass spectrometry allows the characterization and quantitative measurement of these post-translational modifications under different biological conditions. When used in conjunction with measurements of global protein abundance and subcellular localization, greater depth of understanding about cellular response, both to and upon the environment, will emerge.

Emergent work in our lab includes application of proteomics to microbial communities. New advances in both separations and instrumentation resolution have allowed characterization of the protein expression patterns of the microbes within these communities, thus furthering the understanding of how these microbes interact with their environment. For example, we have applied the AMT tag approach to *Pelagibacter ubique* (a.k.a. "SAR11"), perhaps the most abundant microbe in sea-water communities. The data are revealing the adaptive strategies that enable these alphaproteobacteria to recycle carbon efficiently throughout the oceans.

We are also building on this work to include other microbes such as *Caluobacter crescentus* focusing on the determination of important pathways for bioenergy production of both cultured and yet undiscovered or uncultured organisms open the potential for the increase in production of biofuels and the mitigation of the use of these fuels in the environment.

74 ^{GT}L

Functional Genomic Analysis of Current Production in High Power Density Microbial Fuel Cells

Kelly P. Nevin* (knevin@microbio.umass.edu), Sean F. Covalla, Jessica P. Johnson, Trevor L. Woodard, Raymond DiDonato Jr., Kim K. Hixson, Mary Lipton, and **Derek R. Lovley**

Department of Microbiology, University of Massachusetts, Amherst, Massachusetts

Project Goals: The overall purpose of this project is to develop experimental and computational tools to predictively model the behavior of complex microbial communities involved in microbial processes of interest to the Department of Energy. The five year goal is to deliver in silico models that can predict the behavior of two microbial communities of direct relevance to Department of Energy interests: 1) the microbial community responsible for in situ bioremediation of uranium in contaminated subsurface environments; and 2) the microbial community capable of harvesting electricity from waste organic matter and renewable biomass. The research in this abstract summarizes research under Subproject III. The purpose of Subproject III is to define the function of genes of unknown function involved in environmentally significant aspects of the physiology of *Geobacter* species.

Recent advances in the engineering of microbial fuel cells has greatly increased their power output, suggesting that expanded applications, such as powering mobile electronics and large-scale conversion of wastes to electricity, may be feasible goals. Furthermore, although it has widely been considered that mixed microbial communities are required for maximal power output of microbial fuel cells, more recent studies have demonstrated that pure cultures of *Geobacter sulfurreducens* can produce power densities equivalent to those observed with mixed communities. Surprisingly, *G. sulfurreducens* forms thick biofilms on the anodes of such systems and there is a direct correlation between the amount of biomass and current, suggesting that cells at substantial distance from the anode are as effective in contributing to current production as cells in close association with the anode surface. This contrasts with the concept, derived from studies of low power density microbial fuel cells, that cells must be in intimate contact with the anode in order to significantly contribute to power production.

In order to understand the mechanisms for long-range electron transfer to anodes in high power density microbial fuel cells, functional genomic studies of *G. sulfurreducens* growing in flow-through fuel cells are being conducted. For example, a series of deletion mutants which have deficiencies in Fe(III) reduction were evaluated for their ability to produce high levels of current. One gene of particular interest was *pilA*, which encodes the structural pilin protein. Previous studies demonstrated that deleting *pilA* prevents pilin production and Fe(III) oxide reduction. The pilin appear to be electrically conductive and these 'microbial nanowires' are proposed to be the final conduit for electron transfer between the cell and the oxides. Previous studies demonstrated that the *pilA*-deficient mutant produced power as well as wild-type cells in low power density microbial fuel cells. However, in high power density systems current was less than 10% of that observed for wild-type and there has been no adaptation for increased power production in long-term incubations. Expressing *pilA* in *trans* restored current production. Quantitative PCR analysis of *pilA* transcript levels demonstrated that expression of *pilA* increased as current levels, and hence the thickness of the biofilm, increased. These results demonstrate that pili are required for high density current production and suggest that the pilin contribute to long-range electron transfer through the anode biofilm. Pili might also contribute to the structure of the anode biofilm. However, a whole genome comparison of gene

transcript levels with DNA microarrays demonstrated that *pilA* and the pseudopilin gene just downstream of *pilA* were the most highly upregulated genes in current-producing biofilms compared to equally thick biofilms grown on the same graphite surface but with fumarate as the electron acceptor. This indicates an increased need for pilin for electricity production over any potential structural role in the biofilm.

OxpG, is a putative pseudopilin in *G. sulfurreducens* which is part of a type II secretion system necessary for export of proteins essential for Fe(III) oxide reduction to the outer membrane. An *oxpG*-deficient mutant was severely limited in its ability to produce electricity, but produced pili. This suggests that there are outer-membrane proteins other than pili that are essential for current production. Mutants deficient in proteins that are known to be secreted by the type II system are now being evaluated for their capacity for electricity production.

Several outer-membrane *c*-type cytochromes also appear to play a role in high density current production. Previous studies have demonstrated that OmcS is important for low-power density current production. Although transcript levels for *omcS* were elevated in low power density fuel cells, this gene was down regulated in high power density fuel cells. In contrast, expression of genes for other outer-membrane *c*-type cytochromes, such as OmcB and OmcE increased at higher power levels. Although deleting the genes for one or two *c*-type cytochromes typically resulted an initial decrease in power production, these mutants eventually adapted to produce power as well as wild type. This suggests that there is some plasticity in the intermediary pathways of electron transfer to anodes, much more than has been observed in studies on Fe(III) oxide reduction. Proteomic and microarray studies of adapted strains are underway in order to better understand how blockages in electron transfer through important outer-membrane *c*-type cytochromes is overcome. Proteomics studies indicate that outer-membrane *c*-type cytochrome *omcB*, pilin, and numerous hypothetical proteins are present in substantially higher quantities in current harvesting biofilms than in biofilms grown using a soluble electron acceptor.

The periplasmic *c*-type cytochrome, PpcA, is one of the most abundant cytochromes in *G. sulfurreducens* and has previously been shown to be a key intermediary in electron transfer in Fe(III) reduction. Expression of *ppcA* was much 84 fold higher in cells producing current than cells reducing soluble Fe(III). Deleting *ppcA* resulted in a substantial lag in current production and even after long adaption periods the mutant only produced 70% as much power as wild-type cells. Global proteomics analysis to evaluate the mechanisms for adaptation to the loss of PpcA are underway.

The factors leading to optimal power production are complex because they are dependent upon proper biofilm formation, fuel consumption, and multifaceted pathways of extracellular electron transfer. This requires appropriate regulation of many genes. Therefore, the effect of disrupting global regulatory systems is under investigation. For example, deleting the gene for the sigma factor, RpoE, resulted in a significant lag in power production, lower maximum power production than wild type, and a biofilm that was less adherent to the anode surface. Evaluation of a number of other regulatory mutants is in progress.

75 ^{GTL}

Genome-Scale Analysis of Adaptive Evolution of *Geobacter* for Improved Metal Reduction and Electricity Production

Zarath Summers,^{1*} Kelly Nevin,¹ Chris Herring,² Richard Glaven,¹ Shelley Haveman,¹ James Elkins,² Bernhard Palsson,² and **Derek Lovley**¹ (dlovley@microbio.umass.edu)

¹Department of Microbiology, University of Massachusetts, Amherst, Massachusetts and ²Department of Bioengineering, University of California, San Diego, California

Project Goals: The overall purpose of this project is to develop experimental and computational tools to predictively model the behavior of complex microbial communities involved in microbial processes of interest to the Department of Energy. The five year goal is to deliver *in silico* models that can predict the behavior of two microbial communities of direct relevance to Department of Energy interests: 1) the microbial community responsible for *in situ* bioremediation of uranium in contaminated subsurface environments; and 2) the microbial community capable of harvesting electricity from waste organic matter and renewable biomass. The research in this abstract summarizes research under Subproject II. The purpose of Subproject II is to describe gene expression of *Geobacteraceae* under environmental conditions that have relevance for *in situ* uranium bioremediation and harvesting electricity from waste organic matter and to understand factors that may alter this gene expression.

Geobacter species are typically the predominant microorganisms in uranium-contaminated subsurface environments undergoing *in situ* uranium bioremediation and on the surface of electrodes harvesting electricity from organic wastes. The conditions that *Geobacter* species face during *in situ* uranium bioremediation and on the surface of energy-harvesting anodes are substantially different than the conditions under which *Geobacter* species have evolved for billions of years. For example, electron donors are generally limiting in most subsurface environments, but during *in situ* uranium bioremediation electron donors are supplied in excess. Furthermore, in some instances the electron donors supplied, such as lactate, are compounds that are not the primary electron donors for growth of *Geobacter* species under natural conditions. Microbial fuel cells represent a novel environment for *Geobacter* species because, as far as is known, there has been no previous evolutionary pressure on microorganisms to produce electricity. These considerations suggest that long-term exposure of *Geobacter* species to the environmental conditions that prevail during *in situ* uranium bioremediation and in microbial fuel cells will select for changes in gene expression, and possibly beneficial mutations, that will favor growth under these artificially imposed conditions. If so, this could improve the rate and extent of *in situ* uranium bioremediation and enhance power output from microbial fuel cells.

Recent studies with *E. coli* have demonstrated that microorganisms can rapidly evolve when subjected to new environmental constraints. For example, *E. coli* K-12 initially grew much slower on glycerol than predicted from genome-based *in silico* modeling, but growth rates progressively increased toward the predicted optimum as the organism was repeatedly transferred in a minimal medium containing glycerol as the sole carbon and energy source. It has previously been difficult to link such phenotypic variation directly to changes in genotype because subtle differences (i.e. SNPs) could not readily be detected. However, microarray-based comparative genome sequencing revealed that there were a number of spontaneous mutations including SNPs, indels, duplications, and large-scale rearrangements during the adaptation of *E. coli* to faster growth on glycerol. Mutations in genes encoding glycerol kinase (*glpK*) and large subunits of the RNA polymerase (*rpoC* and *rpoB*) resulted in the greatest fitness for growth on glycerol and these mutations were rapidly fixed in the

evolving populations. Evolved growth phenotypes could be reconstructed in the wild-type strain by introducing the experimentally determined mutations via site-directed mutagenesis. These studies demonstrated that genome-wide perturbations can be identified during laboratory-scale evolution studies and that causal mutations directly linking evolved genotypes to the resulting phenotypes can be determined. These methods developed with *E. coli* can now be applied to adaptive evolution studies with *Geobacter* species.

In order to understand how *Geobacter* species might evolve during *in situ* uranium bioremediation or harvesting electricity from waste organic matter, adaptive evolution studies were initiated with *Geobacter sulfurreducens*. For example, to determine if *G. sulfurreducens* could be adapted for more rapid extracellular electron transfer, it was repeatedly transferred with Fe(III) oxide as the electron acceptor under conditions which favored rapid growth on Fe(III) oxide. A strain was developed which can transfer electrons to Fe(III) and Mn(IV) oxides 10 times faster than the unadapted strain. Analysis of gene transcript levels with whole-genome DNA microarrays demonstrated that the evolved strain had higher levels of expression of genes for proteins thought to be involved in electron transfer to Fe(III) oxides, such as *c*-type cytochromes, and pili. Transcripts of genes encoding transport proteins, central metabolism enzymes, and several hypothetical proteins were also higher in the adapted strain. Surprisingly, when the gene for the outer-membrane *c*-type cytochrome, OmcS, was deleted in the adapted strain this had no impact on Fe(III) oxide reduction whereas OmcS is required for Fe(III) oxide reduction in the unadapted strain. This suggests that the adapted strain of *G. sulfurreducens* has significant changes in its extracellular electron transport chain. One key mutation in the adapted strain has already been detected and comparative genome sequencing is in progress.

Other adaptive evolution studies are underway. For example, the lower the potential that microorganisms transfer electrons to the anodes of microbial fuel cells, the greater the power production. A strain of *G. sulfurreducens* has been adapted to transfer electrons to a fuel cell anode at much lower potential than the wild-type strain. Lactate is a convenient electron donor source to add to the subsurface in order to promote *in situ* uranium bioremediation, but lactate is not a common electron donor in natural anaerobic sedimentary environments. Therefore, studies have been initiated to elucidate how *G. sulfurreducens* adapts for enhanced lactate utilization with continual exposure to excess lactate as the sole electron donor. Additional selective pressures that have relevance to *in situ* uranium bioremediation and/or optimizing power output of microbial fuel cells are also in progress.

76 ^{GTL}

Proteomic Profiling of the *Caulobacter crescentus* Cell Cycle and Starvation Response

Esteban Toro^{1*} (etoro@stanford.edu), Leticia Britos,¹ Samuel O. Purvine,² Mary S. Lipton,² Tom Taverner,¹ Feng Yang,² **Harley H. McAdams**,¹ Richard D. Smith,² and Lucy Shapiro^{1*}

¹Department of Developmental Biology, Stanford University, Stanford, California and ²Pacific Northwest National Laboratory, Richland, Washington

Project Goals: Identification of the overall regulatory and metabolic networks in *Caulobacter crescentus*, largely through gene expression microarray assays and bioinformatic analysis

The gram-negative bacterium, *Caulobacter crescentus*, is closely related to economically important plant symbionts (e.g. *Rhizobium spp.*) and human pathogens (e.g. *Rickettsia spp.*), is ubiquitous, innocuous, and easily manipulated with standard genetic methods. *Caulobacter* cells are asymmetric, dividing into two distinct cell types that can be easily differentiated using light microscopy. Most importantly, *Caulobacter* cultures can be synchronized with a simple procedure, allowing researchers to investigate the modulation of cellular processes during the cell cycle and determine the detailed molecular mechanisms governing cellular operations. We have used liquid chromatography coupled to mass spectrometry (LC-MS and LC-MS/MS) to: i) directly determine the relative levels of all *Caulobacter* proteins during the cell cycle; ii) compare the proteomic profile of exponentially growing cells to stationary cells and cells being starved for carbon; and iii) quantitate absolute levels of key regulatory proteins.

Using 2-dimensional sample fractionation, with strong cation exchange (SCX) and reverse phase liquid chromatography coupled to tandem MS analysis (LC-MS/MS) we have confidently identified 3174 distinct *Caulobacter* proteins (accounting for about 84% of *C. crescentus* predicted genes) in exponentially growing cells. These proteins included inner and outer membrane proteins as well as proteins with extreme pIs that are difficult to resolve using gel electrophoresis-based methodologies. This number includes only proteins identified using a conservative methodology which requires at least two unique peptides and unambiguous MS/MS characterization before a protein hit is called. Comparing our peptide probability results to independent predictions from PeptideProphet shows our false positive rate to be below 5%.

Taking advantage of the ability to synchronize *Caulobacter* cultures, we performed global proteomics measurements as the cells progressed through the cell cycle. Furthermore, in order to obtain a preliminary view of how information from the environment is processed, we have done proteomic experiments to measure *Caulobacter's* response to carbon starvation and stationary phase. Finally, we describe initial results on absolute quantitation of key regulatory proteins throughout the *Caulobacter* cell cycle.

77 ^{GTL}

Quantitative Shotgun Proteomics with ProRata: Application to Anaerobic Aromatic Degradation in *Rhodopseudomonas palustris*

C. Pan,^{1*} G. Kora,¹ Y. Oda,² D. Pelletier,¹ N. C. VerBerkmoes,¹ W. H. McDonald,¹ G. Hurst,¹ C. S. Harwood,² R. L. Hettich¹ (hettichrl@ornl.gov), and **N. F. Samatova**¹ (samatovan@ornl.gov)

¹Oak Ridge National Laboratory, Oak Ridge, Tennessee and ²University of Washington, Seattle, Washington

Project Goals: Quantitative shotgun proteomics has recently emerged as a high throughput technique for measuring the relative abundances of thousands of proteins between two cellular conditions. The project addressed the following challenges: 1) accurate estimation of protein abundance ratios from noisy mass spectral data; (2) extraction of reliable information from a biological comparison. It then aims to characterize the catabolism of p-coumarate in *Rhodopseudomonas palustris*.

Organisms often respond to environmental or physiological stimuli by adjusting the type and abundance of proteins in their cells. Measurement of the relative abundances of proteins in treatment cells subjected to stimuli, compared to that in the reference cells, provides valuable insights about protein function and regulation. Quantitative shotgun proteomics has recently emerged as a high throughput

technique for measuring the relative abundances of thousands of proteins between two cellular conditions. The reference and treatment proteomes are labeled with different stable isotope tags and then mixed in equivalent amounts. In such a proteome mixture, each protein has two mass-different isotopic variants: the light isotopologue and the heavy isotopologue. The proteome mixture is digested and then analyzed with liquid chromatography–tandem mass spectrometry (LC–MS/MS). There are two folds of informatics challenges in quantitative proteomics: 1) accurate estimation of protein abundance ratios from noisy mass spectral data; (2) extraction of reliable information from a biological comparison.

ProRata: A quantitative proteomics program for accurate protein abundance ratio estimation with confidence interval evaluation [1, 2].

The abundance ratio between the light and heavy isotopologues of an isotopically labeled peptide can be estimated from their selected ion chromatograms. However, quantitative shotgun proteomics measurements yield selected ion chromatograms at highly variable signal-to-noise ratios for tens of thousands of peptides. This challenge calls for algorithms that not only robustly estimate the abundance ratios of different peptides but also rigorously score each abundance ratio for the expected estimation bias and variability. Scoring of the abundance ratios, much like scoring of sequence assignment for tandem mass spectra by peptide identification algorithms, enables filtering of unreliable peptide quantification and use of formal statistical inference in the subsequent protein abundance ratio estimation. In this study, a parallel paired covariance algorithm is used for robust peak detection in selected ion chromatograms. A peak profile is generated for each peptide, which is a scatter-plot of ion intensities measured for the two isotopologues within their chromatographic peaks. Principal component analysis of the peak profile is proposed to estimate the peptide abundance ratio and to score the estimation with the signal-to-noise ratio of the peak profile (profile signal-to-noise ratio). We demonstrate that the profile signal-to-noise ratio is inversely correlated with the variability and bias of peptide abundance ratio estimation. Then, a profile likelihood algorithm is proposed to infer the abundance ratios of proteins from the abundance ratios of isotopically labeled peptides. Given multiple quantified peptides for a protein, the profile likelihood algorithm probabilistically weighs the peptide abundance ratios by their inferred estimation variability, accounts for their expected estimation bias, and suppresses contribution from outliers. This algorithm yields maximum likelihood point estimation and profile likelihood confidence interval estimation of protein abundance ratios. This point estimator is more accurate than an estimator based on the average of peptide abundance ratios. The confidence interval estimation provides an “error bar” for each protein abundance ratio that reflects its estimation precision and statistical uncertainty. The accuracy of the point estimation and the precision and confidence level of the interval estimation were benchmarked with standard mixtures of isotopically labeled proteomes. The parallel paired covariance algorithm, the principal component analysis algorithm and the profile likelihood algorithm were integrated into a quantitative proteomics program, called ProRata, freely available at www.MSProRata.org.

Characterization of Anaerobic Catabolism of *p*-Coumarate in *Rhodopseudomonas palustris* by Integrating Transcriptomics and Quantitative Proteomics [3].

In this study, the pathway for anaerobic catabolism of *p*-coumarate by a model bacterium, *Rhodopseudomonas palustris*, was characterized by comparing its gene expression profile under *p*-coumarate growth against those under succinate and benzoate growth. Gene expression was quantified at the mRNA level with transcriptomics and at the protein level with quantitative proteomics using ¹⁵N metabolic labeling. Both -omics measurements were critical, since the transcriptomics provided near full genome coverage of gene expression profiles and the quantitative proteomics surveyed the expression activities of over 1,500 genes at the protein level. The integrated gene expression data are consistent with the proposal that *p*-coumarate is converted to benzoyl-CoA, which is then degraded

via a known aromatic ring reduction pathway. For the metabolism of *p*-coumarate to benzoyl-CoA, two alternative routes, a β -oxidation route and a non- β -oxidation route, are possible. Based on the integrated gene expression data, we suggest that the anaerobic catabolism of *p*-coumarate likely proceeds through the non- β -oxidation route in *R. palustris*. A putative gene was proposed for every step in the non- β -oxidation route.

Research sponsored by the DOE Genomics:GTL Program.

References

1. C. Pan, G. Kora, D.L. Tabb, D.A. Pelletier, W.H. McDonald, G.B. Hurst, R.L. Hettich, and N.F. Samatova, Robust estimation of peptide abundance ratios and rigorous scoring of their variability and bias in quantitative shotgun proteomics. *Analytical Chemistry*, 2006. **78**(20): p. 7110-7120.
2. C. Pan, G. Kora, W.H. McDonald, D.L. Tabb, N.C. VerBerkmoes, G.B. Hurst, D.A. Pelletier, N.F. Samatova, and R.L. Hettich, ProRata: A quantitative proteomics program for accurate protein abundance ratio estimation with confidence interval evaluation. *Analytical Chemistry*, 2006. **78**(20): p. 7121-7131.
3. C. Pan, Y. Oda, D.A. Pelletier, B. Zhang, P.K. Lankford, N.F. Samatova, C.S. Harwood, R.L. Hettich. Characterization of Anaerobic Catabolism of *p*-Coumarate in *Rhodopseudomonas palustris* by Integrating Transcriptomics and Quantitative Proteomics. (Submitted)

78 ^{GTL}

From Genome to Metabolome: Correlating a System-Wide Response to Environmental Adaptation in a Hyperthermophile

Sunia A. Trauger^{1*} (strauger@scripps.edu), Ewa Kalisiak,¹ Jarek Kalisiak,² Hiro Morita,¹ Angeli Lal Menon,³ Michael V. Weinberg,³ Farris L. Poole,³ Michael W. W. Adams,³ and **Gary Siuzdak**¹

¹Center for Mass Spectrometry and the ²Department of Chemistry, The Scripps Research Institute, La Jolla, California; and ³Department of Biochemistry and Molecular Biology, University of Georgia, Athens, Georgia

Project Goals: Watson and Crick's discovery united structure and genetics in a very powerful way. With the combined effort of the MAGGIE investigators, we seek to achieve a similar type of unification for gene sequence and cell biology. MAGGIE investigators have developed model microbial systems and technologies suitable to derive a prototypical multi-level map of protein modifications and interactions. We believe that the requested level of effort and funding is merited as we are working to solve an extremely challenging recognition problem. For example, it's not obvious from the surfaces of proteins how they interact specifically and how they will functionally switch conformational states. Yet, given the resources of MAGGIE and the establishment of coupled gene-biology cycles for the microbial systems investigated by this Program, we can experimentally define these modifications and interactions promoting protein functions. Long term, MAGGIE seeks to test the underlying hypothesis that the architecture of proteins and their complexes encodes in part how individual steps of pathways are coupled to each other to form coherent pathways; and how these pathways interact without disruptive interference. This idea is built upon both experimental results from the Program members and considerations rooted in first principles about the nature of macromolecular interactions.

Hyperthermophiles represent a unique group of prokaryotic microorganisms that optimally grow at temperatures exceeding those normally observed for other organisms (at or above 80°C).¹ They are

biochemically and evolutionarily unique organisms that have adapted to the challenges of molecular and structural stability of their higher temperature habitats. Important biomolecules such as DNA, RNA and proteins can undergo deleterious changes at higher temperatures for most organisms.² However, these deleterious effects are largely suppressed for hyperthermophiles. This suggests that the unique evolutionary history and adaptation to their geo-thermal habitats has led to the development of novel mechanisms of DNA and protein stabilization. Hyperthermophiles do not perish at cooler temperatures, but undergo an adaptation response that allows them to grow in a sustained manner after an initial cold shock.³ To better understand the biochemical changes in these unique organisms grown at their optimal 95°C and those that have adapted to 72°C, we undertook a study of the molecular changes through a systems-wide mass spectrometry based approach that involves comparative metabolite and proteomic profiling at the two environmental conditions followed by the identification of molecules that differentiate the two cell populations of *Pyrococcus furiosus*.

Metabolites were extracted from the soluble fraction of cells grown at 95°C and 72°C using cold acetone precipitation. The protein pellet from this extraction procedure was reserved for separate proteomic analysis. The extract was analyzed using LC-MS on an Agilent MSD-TOF mass spectrometer that routinely yields accurate masses (<5 ppm). LC-MS data from cells grown at both temperatures was analyzed using the XCMS program developed in the Siuzdak laboratory.⁴ This software package allows for the non-linear alignment of chromatograms and the identification of significant differences between multiple samples. Using this approach, many features differentiated the two samples. Most metabolites were down regulated in samples grown at 72°C compared to 95°C. These included amino acids like arginine, phenylalanine, and leucine. However, there were some metabolites that actually underwent an up-regulation. One of these could be identified through accurate mass measurement and tandem mass. A previously unidentified molecule was discovered to be up-regulated during cold adaptation and was finally identified as N-acetylthermospermine. Thermospermine has previously been identified in other organism in Archaea.⁵ The identification of these polyamines as being up-regulated during cold adaptation response is highly significant since these are known to be involved in DNA/RNA stabilization in other organisms.

In order to better understand the overall proteomic changes as *Pyrococcus furiosus* adapts to the colder 72 °C environment, a proteomic profiling experiment was designed. Trypsin digests were prepared from the protein pellets reserved from the metabolite profiling experiment. Protein concentrations were measured using a Bradford assay and the concentrations were normalized for two samples after re-dissolving the pellets. Since the *Pyrococcus* proteome has low cysteine content, thiol specific quantitative proteomic techniques such as isotope coded affinity tags (ICAT) were not suitable. We decided to use a spectral counting approach to relative protein quantitation using the ESI-TOF, and pursue protein identification using tandem mass spectrometry using a linear ion trap. Nano-LC-MS analyses were performed on the ESI-TOF mass spectrometer, as well as on the Finnigan LTQ using a nano-LC system on a mobile cart to minimize any differences in the chromatographic retention times. LC-MS/MS data on the ion trap was used for protein identification using Mascot (Matrix Science). This resulted in the identification of over 200 proteins. XCMS analysis was used to perform the quantitative proteomic analysis, as well as to identify the major differences between the samples grown under the two different conditions. The top 100 peptide ions observed to undergo significant change were identified through accurate mass and tandem MS experiments performed on the LTQ within narrow windows of retention time. Most proteins showed a down-regulation trend as the organism adapts to 72°C, which is consistent with the gene expression data on *P. furiosus* grown at 95°C and 72°C.⁵ However, some proteins like ABC transporter showed an up-regulation.

The identification of metabolites such as spermidine, raises the question of which enzymes are involved with the regulation of these metabolites in hyperthermophiles. An experiment was designed

to do a targeted proteomic analysis using immobilized spermidine as a probe. Affinity columns were prepared by covalently binding spermidine to beads (Microlink kit from Pierce). Control columns were also prepared without spermidine. These were incubated overnight with the cytoplasmic fraction from the sample grown at 72°C. After several washes to remove non-specific binding, the nano-LC-MS/MS analysis of the eluted protein samples after trypsin digestion resulted in the identification of over a hundred proteins when compared to the control. Proteins identified included an putative acetyl transferase, archeal histones and three proteins previously identified as being up-regulated in the DNA microarray experiments.⁵

In this study, we demonstrate a comprehensive approach to metabolite and proteomic profiling using mass spectrometry that not only allows the identification of a novel metabolite, but also the characterization of changes in the proteome. Finally the metabolite immobilization and proteomics approach allows us to discover protein candidates that may help regulate key metabolites in the cold adaptation response.

References

1. Stteter K (1996). Hyperthermophilic prokaryotes. *FEMS Micro. Biol. Rev.* 18: 149-158.
2. Marmur J., and Doty, P. (1962) Determination of the base composition of deoxyribonucleic acid from its thermal denaturation temperature. *J. Mol Biol* 5, 109-118.
3. Dennis W. Grogan (1998). Hyperthermophiles and the problem of DNA instability. *Molecular Microbiology*. 28(6),1043-1049.
4. Smith C.A., Want E.J., O'Maille G., Abagyan R., Siuzdak G. (2006). XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching and identification. *Analytical Chemistry* 78, 779-787.
5. Weinberg, M. V., Schut, G. J., Brehm, S., Datta, S., Adams M. W.W. (2005), Cold shock of a hyperthermophilic archaeon: *Pyrococcus furiosus* exhibits multiple responses to a suboptimal growth temperature with a key role for membrane-bound glycoprotein. *J. Bacteriology* 187 (1), 336-348.

79 GTL

High Throughput Comprehensive and Quantitative Microbial and Community Proteomics

Richard D. Smith* (rds@pnl.gov), Joshua N. Adkins, David J. Anderson, Kenneth J. Auberry, Mikhail E. Belov, Stephen J. Callister, Therese R.W. Clauss, Jim K. Fredrickson, Xiuxia Du, Kim K. Hixson, Navdeep Jaitly, Gary R. Kiebel, Mary S. Lipton, Eric A. Livesay, Anoop Mayampurath, Matthew E. Monroe, Ronald J. Moore, Heather M. Mottaz, Carrie D. Nicora, Angela D. Norbeck, Daniel J. Orton, Ljiljana Paša-Tolić, Kostantinos Petritis, David C. Prior, Samuel O. Purvine, Yufeng Shen, Anil K. Shukla, Aleksey V. Tolmachev, Nikola Tolić, Harold R. Udseth, Rui Zhang, and Rui Zhao

Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington

Project Goals: Our primary goals are to significantly increase analysis throughput and data quality, and capitalize on these advances to enable the study of increasingly complex microbial communities.

Significance: Capabilities for quantitative proteomics measurements of steadily increasing throughput and quality have been implemented and are being applied to studies of diverse microbial systems, and more increasingly to microbial communities.

With recent advances in whole genome sequencing for a growing number of organisms, biological research is increasingly incorporating higher-level “systems” perspectives and approaches. For example, in nature, microbial cells rarely exist as individual colonies, but interact with other microbes in a community and with their environment, thus creating an ecosystem. The challenges of studying these higher-levels systems, such as a microbial ecosystem, are effectively open-ended due to the complexity of microbial communities, the number of possible interactions, and the technology that allows us to more completely observe complex systems. Advancing a systems-level understanding of microbial and other biological research is at the heart of the DOE Genomics:GTL program.

One aspect pertinent to a systems-level understanding is the ability to quantitatively measure the array of proteins (i.e., the proteome) under stable and perturbed conditions and from naturally occurring microbial communities. Among the challenges associated with making useful comprehensive proteomic measurements are identifying and quantifying large sets of proteins whose relative abundances span many orders of magnitude. Additionally, these proteins may vary broadly in chemical and physical properties, have transient and low levels of modifications, and be subject to endogenous proteolytic processing. Ultimately, such measurements and the resulting insight into biochemical processes are expected to enable development of predictive computational models that could profoundly affect environmental clean-up, understandings related to climate, and energy production by e.g., providing a more solid basis for mitigating the impacts of energy production-related activities on the environment.

A “prototype high throughput production” lab established in FY 2002 was an early step towards implementing higher throughput proteomics measurements. Operations within this lab are distinct from technology development efforts, both in laboratory space and staffing. This step was instituted in recognition of the different staff “mind sets” required for success in these different areas, as well as to allow “periodic upgrades” of the technology platform in a manner that does not significantly impact its production operation. The result has been faster implementation of technology advances and more robust automation of technologies that improve overall effectiveness.

The biological applications of the technology and associated activities are the subject of a separate, but interrelated project (J. K. Fredrickson, PI), involving studies of a number of individual microbial systems (e.g., *Shewanella oneidensis*, *Geobacter sulfurreducens*, *Rhodobacter sphaeroides*, *Caulobacter crescentus*) and communities (e.g., SAR 11 marine community) in collaboration with leading experts. These studies have demonstrated the capability for automated high-confidence protein identifications, broad proteome coverage, and for exploiting both stable isotope labeling and label-free methods to obtain high precision in protein abundance measurements.

With a paradigm established for high throughput proteomic measurements, our primary goals are to significantly increase analysis throughput and data quality, and capitalize on these advances to enable the study of increasingly complex microbial communities. A significant challenge is how to maximize the information content derived from large and complex data sets to obtain improved understanding of biological systems. Thus, a key component of our program involves developing the informatics tools needed to quantify and define the quality of data, as well as the tools to make the results broadly available and understandable to the researchers. Efforts currently in progress aim to:

- Significantly increase the overall data production by more than an order of magnitude in conjunction with increased data quality, providing data that are quantitative and have statistically-based measures of quality.
- Extend the application to an increasing number of different kinds of post-translation modifications.

- Apply improved data quality and improved sample processing for high throughput measurements of increasingly complex microbial communities.
- Provide the infrastructure and informatics tools required to efficiently manage, use, and disseminate large quantities of data generated by GTL “users.”

This presentation will highlight the advances in providing high quality data with statistically-founded measures of quality, while providing increased measurement throughput. The advances will be illustrated in the context of applications to microbes and microbial communities of interest to the GTL program.

Acknowledgements: This research is supported by the Office of Biological and Environmental Research of the U.S. Department of Energy. Pacific Northwest National Laboratory is operated for the U.S. Department of Energy by Battelle Memorial Institute through Contract No. DE-AC05-76RLO 1830.

80 ^{GT}L

Exploring the Genome and Proteome of *Desulfitobacterium hafniense* DCB-2 for its Protein Complexes Involved in the Reduction of Selenium and Iron

Christina Harzman,¹ Christi Hemming,¹ Sang-Hoon Kim,¹ David DeWitt,¹ John Davis,² Rachel Udelhoven,³ Kaitlin Duschene,³ Joan B. Broderick,³ **James M. Tiedje**,¹ Terence L. Marsh^{1*}
(marsht@msu.edu)

¹Michigan State University, East Lansing, Michigan; ²Columbus State University, Columbus, Georgia; and ³Montana State University, Bozeman, Montana

Project Goals: The goal of our investigations into the cell biology of *Desulfitobacterium hafniense* is to identify the functional genomics/proteomics of metal reduction and determine the chemistry of relevant enzymatic activities. To this end we have investigated the regulation of growth under environmental conditions and have identified the genes and proteins up and down-regulated when metals were used as terminal electron acceptors. The identification of putative pathways and proteins through gene arrays and proteomics under metal reducing conditions will be confirmed using genetic knock-outs and putative activities explored through cloning and overexpression. In this way, the response of *D. hafniense* to its environment will be better understood and approaches to successfully employ it as an ally in bioremediation can be designed.

Desulfitobacterium hafniense is an anaerobic, low GC Gram-positive, spore-forming rod with remediation capabilities that include chlororespiration and metal reduction. The latter has been of interest to us as we have previously reported that the metal reduction capabilities of the organism include Fe(III), Cu(II), Co(III), Se(VI), and U(VI). We have also demonstrated that *D. hafniense* can form biofilms under metal-reducing and fermentative conditions. Our investigations of selenium reduction have revealed morphological changes to the cell, especially on the surface. In response to 1 mM Se(VI), *D. hafniense* cells elongate and form vesicular blebs on the cell's surface. The vesicles appear to eventually bud off of the cell and can be seen as spherical entities in the milieu with a diameter of approximately 0.2 μ m (See Fig.1). These vesicles are bound by both membrane and cell wall and contain high concentrations of selenium as judged by energy dispersive x-ray spectroscopy. Hence, we have postulated that these vesicles detoxify through sequestration of the selenium.

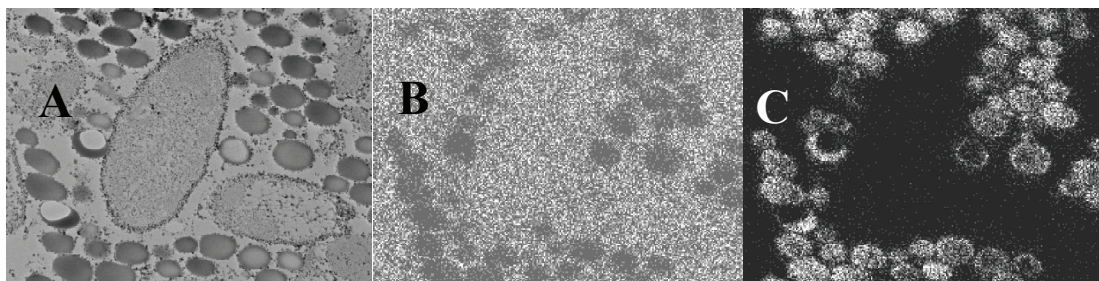


Figure 1. *D. hafniense* grown fermentatively in the presence of 1 mM selenium.

A. negatively stained transmission EM showing two cells surrounded by smaller vesicles, B. same field scanned for carbon; note low carbon content of vesicles, C. same field scanned for selenium;

Genomics

The genome of *D. hafniense* has been sequenced and we have constructed a gene array containing 4,715 targeted ORFs. Towards a more detailed understanding of the components required for metal reduction, we have performed transcriptome and proteome analysis of *D. hafniense* growing fermentatively with and without 1 mM selenium. In the presence of selenium, 27 putative operons were up-regulated and 34 were down-regulated. A total of 304 genes were up-regulated and 376 down-regulated at least 2 fold. Notable up-regulated genes coded for antitoxin RelB, transcriptional regulators, radical SAM, ABC transporters, ferredoxins, inorganic ion transport, cytochromes, DMSO reductases, heavy metal translocation, and two MinD-like proteins similar to cell septation protein MinD. Down-regulated genes included genes for NADH dehydrogenase subunits, coenzyme metabolism, flagella biosynthesis, ABC transporters, amino acid transport and metabolism. In addition, many hypothetical genes were down-regulated severely.

The formation of vesicles in response to selenium is consistent with altered regulation of *minD*, which is involved in cell septation and minicell formation. Attempts are being made to overproduce the MinD-like protein in *D. hafniense* in order to test its effect on vesicle formation with and without Se(VI). The machinery capable of reducing Se(VI) is being examined by constructing knock-out mutants defective in the neighboring [Fe-S] protein genes that are likely co-transcribed with the *minD*-like genes. Finally, a lipoprotein transporter (LolCDE) that is usually found in Gram-negative bacteria has been detected in *D. hafniense* as transcriptionally active with selenium. This is being investigated with gene disruption approaches.

In contrast to fermentative growth, 72 operons were up-regulated and 74 were down-regulated when Fe(III) was used as the terminal electron acceptor. A total of 678 genes were up-regulated and 643 genes were down-regulated relative to fermentative growth. Up-regulated groups include: lactate transport, ferredoxins, flavoproteins, cytochromes, radical SAM, DMSO reductases, nitrogenases, cell envelope biogenesis, heat shock proteins, heavy metal translocation, antitoxin genes (*relB*), and drug resistance genes.

Proteomics

A MarR transcription factor, a histidine kinase, and multiple proteins from each of three operons have been identified by proteomics analyses as elevated in *D. hafniense* cells grown in 1 mM Se(VI), and are proposed to be involved in selenium respiration by this organism. The MarR transcription factor and histidine kinase were elevated 65- and 25-fold respectively. The role of the MarR during

selenium reduction is not known, but histidine kinases have been implicated in the elevated expression of reductive dehalogenases in *D. hafniense*. A nickel dependent hydrogen oxidase subunit and b-type cytochrome subunit belonging to an operon similar to that shown previously to provide reducing equivalents for dehalogenation in *D. dehalogens* were also elevated. Four proteins in a DMSO reductase operon including a periplasmic (twin-arginine signal) anaerobic selenocysteine containing dehydrogenase were also found to be elevated. Four genes in a third operon, two of which code for the molybdenum cofactors necessary for DMSO reductase activity were elevated on average 5 fold at the protein level and 5-16 fold by microarray analysis. These studies suggest that respiratory reduction of selenium occurs via a DMSO reductase, using electrons provided by hydrogen oxidation, and provide an explanation for the unique function of one of the more than 50 DMSO operons in *D. hafniense*.

***In vitro* biochemical analysis**

Proteins identified by proteomics and genomics analysis as relevant to metal reduction by *D. hafniense* are being cloned for expression in *E. coli*, in order to isolate sufficient protein for detailed functional and biochemical analysis. Initial targets include DMSO reductases and sulfite reductases, as these are in some cases significantly up-regulated and are likely catalysts for selenate reduction. Other targets include several of the putative radical-SAM proteins observed to be up-regulated; some of these are likely involved in biosynthesis of the molybdopterin cofactors found in enzymes such as DMSO reductase and sulfite reductase. Target genes have been amplified from *D. hafniense* genomic DNA by PCR, and have been inserted into expression vectors using the Gateway system. Preliminary characterization of the heterologously expressed proteins will be presented.

Section 2

Metabolic Network Experimentation and Modeling

81

 MEWG

Improving the Production of Biotherapeutics using Metabolic Engineering

M. Bauman,^{1*} J. Jones,¹ S. Krag,³ V. Ciccarone,⁴ D. Judd,⁵ S. Gorfien,⁵ Y.C. Lee,² N. Tomiya,² and **M. Betenbaugh^{1*}** (beten@jhu.edu)

¹Departments of Chemical and Biomolecular Engineering, ²Biology, and ³Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland; ⁴MacroGenics Corporation, Rockville, Maryland; and ⁵Gibco, Invitrogen Corporation, Grand Island, New York

Project Goals: We are investigating the metabolic pathways involved in protein N-glycosylation. Glycoproteins that are expressed in mammalian cell culture systems often have variable N-glycosylation and we are evaluating the potential causes of this phenomenon. Using metabolic engineering strategies, we plan to overcome any bottlenecks in the pathways that may lead to inefficient or improper N-glycosylation, and therefore to improve the yield and quality of biotherapeutics that can be produced in mammalian cell culture systems.

Many of the valuable commercial biotherapeutic products, such as monoclonal antibodies, growth factors, hormones and clotting factors, are secreted glycoproteins. These biotherapeutics are often produced in mammalian cell expression systems and are made up of oligosaccharide chains attached to polypeptides at specific amino acid residues. The addition of these oligosaccharides to the proteins occurs through a post-translational modification called N-glycosylation. The number, type, and location of the oligosaccharides (glycans) on the protein can affect key biochemical properties of the biotherapeutic, including its clearance rate, immunogenicity, bioactivity, solubility, and stability against proteolysis. Unfortunately, when these therapeutic products are produced by over-expression in mammalian and non-mammalian hosts, the glycosylation processing can generate products with highly variable glycosylation patterns. This N-glycan variability limits the yield and affects the quality of the target secreted glycoproteins and therefore can significantly affect the value of biotherapeutic products.

In an attempt to overcome the problems associated with variable N-glycosylation, we are investigating the metabolic pathways involved and evaluating potential causes of this phenomenon. Specifically, we are examining the N-linked glycosylation process that involves the transfer of a pre-formed oligosaccharide onto an acceptor Asparagine residue on a nascent polypeptide in the lumen of the endoplasmic reticulum (ER). Unfortunately, these N-glycan acceptor sites are not always fully occupied, leading to site occupancy heterogeneity. Using human transferrin (hTf) as a model protein, we have shown that variable site occupancy occurs when over-expressing certain glycoproteins using two different mammalian cell culture systems. Examination of hTf using SDS-PAGE as well as MECC (Micellar Electrokinetic Chromatography) reveals a difference in the molecular weight profile between the intracellular and secreted fractions. Treatment with tunicamycin (a glycosylation inhibitor) abrogates this difference, implying that N-glycosylation is responsible for the size difference. Immunoprecipitation experiments suggest that, instead of being secreted, under-glycosylated hTf interacts with ER chaperone proteins and accumulates intracellularly. We plan to use metabolic engineering strategies to overcome bottlenecks in the N-glycosylation pathway that lead to the formation of under-glycosylated proteins in mammalian cell culture systems.

82 ^{GTL}

Improved Microbial Hydrogen Production by the Engineering of Specific Metabolic Segments of *Escherichia coli*

Zhanmin Fan,¹ Ling Yuan,¹ Yu Wang,² and **Ranjini Chatterjee**^{2*} (rchatterjee@farasis.com)

¹Department of Agronomy, University of Kentucky, Lexington, Kentucky and ²Farasis Energy, Inc., Hayward, California

Project Goals: 1. The development of *Escherichia coli* as a platform microbe for the production two clean fuels, hydrogen and ethanol, and the commodity chemical succinic acid. 2. The construction of a unique high throughput gas chromatography (GC) instrument for detection of hydrogen in micro-scale fermentations.

Hydrogen (H₂) has significant potential as a clean energy source to replace non-renewable and polluting fossil fuels. Biological routes to H₂ production represent environmentally benign processes that utilize renewable raw material derived from agricultural products, and microbes or enzymes as the catalysts for energy generation. Development of *Escherichia coli* (*E. coli*) as a biocatalyst for H₂-production offers solutions to some of the challenges facing biological processes for H₂. An endogenous H₂-evolving activity exists in *E. coli* under fermentative conditions: the formate hydrogenlyase

(FHL) enzyme complex catalyzes the disproportionation of formate to CO_2 and H_2 and is the subject of our research. The production of a functional FHL complex was improved by systematically engineering key regulatory, and amino acid biosynthetic pathways of *E. coli*. The engineered *E. coli* strains produced elevated levels of H_2 compared to the wild-type (WT) strain. The results from this program provide valuable insights into a panel of auxiliary proteins and metabolic pathways that can be engineered to increase H_2 production by microbes.

83

Toward the Automatic Generation of Genome-Scale Metabolic Models in the SEED

Matthew DeJongh^{1*} (dejongh@hope.edu) and Aaron Best²

¹Department of Computer Science and ²Department of Biology, Hope College, Holland, Michigan

Project Goals: We are developing a method for automating the generation of genome-scale metabolic models suitable for analytical techniques such as flux-balance analysis. We are implementing our approach within the SEED, a software environment for comparative genome annotation and analysis (www.theseed.org). Our technology sets the stage for the automatic generation of substantially accurate metabolic reconstructions for over 400 complete genome sequences currently in the SEED.

Current methods for the automatic generation of genome-scale metabolic models focus heavily on genome annotation and preliminary biochemical reaction network assembly, but do not adequately address the process of identifying and filling gaps in the reaction network, and verifying that the network is suitable for systems level analysis. Thus, current methods are only sufficient for generating draft-quality models, and refinement of the reaction network is still largely a manual, labor-intensive process [1].

We have developed a method for automating the generation of genome-scale metabolic models that produces substantially complete reaction networks, suitable for analytical techniques such as flux-balance analysis. Our method partitions the reaction space of central and intermediate metabolism into discrete, interconnected components that can be assembled and verified in isolation from each other, and then integrated and verified at the level of their interconnectivity. We have developed a database of components that are common across organisms, and have created tools for automatically assembling appropriate components for a particular organism based on the metabolic pathways encoded in the organism's genome. This focuses manual efforts on those portions of an organism's metabolism that are not yet represented in the database. We have demonstrated the efficacy of our method by reverse-engineering and automatically regenerating the reaction network from a published genome-scale metabolic model for *Staphylococcus aureus* [2]. Additionally, we have created initial reconstructions of three other published metabolic models (*Escherichia coli* [3], *Helicobacter pylori* [4], and *Lactococcus lactis* [5]) to demonstrate that our approach reduces the manual effort involved in model creation, by building on the common reaction network components already created for the *S. aureus* model. We have implemented our tools and database within the SEED, a software environment for comparative genome annotation and analysis (www.theseed.org) [6].

Our technology sets the stage for the automatic generation of substantially accurate metabolic reconstructions for over 400 complete genome sequences currently in the SEED. With each genome that

is processed using our tools, the database of common components grows to cover more of the diversity of metabolic pathways, further reducing the manual effort involved in generating subsequent genome-scale metabolic models for other sequenced organisms.

Research sponsored by Argonne National Laboratory and Howard Hughes Medical Institute.

References

1. Francke, C., R.J. Siezen, and B. Teusink (2005). "Reconstructing the metabolic network of a bacterium from its genome." *Trends Microbiol.* **13**(11): p. 550-8.
2. Becker, S.A. and B.O. Palsson (2005). "Genome-scale reconstruction of the metabolic network in *Staphylococcus aureus* N315: an initial draft to the two-dimensional annotation." *BMC Microbiol.* **5**(1): p. 8.
3. Reed, J.L., et al. (2003). "An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR)." *Genome Biol.* **4**(9): p. R54.
4. Schilling, C.H., et al. (2002). "Genome-scale metabolic model of *Helicobacter pylori* 26695." *J Bacteriol.* **184**(16): p. 4582-93.
5. Oliveira, A.P., J. Nielsen, and J. Forster (2005). "Modeling *Lactococcus lactis* using a genome-scale flux model." *BMC Microbiol.* **5**: p. 39.
6. Overbeek, R., et al. (2005). "The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes." *Nucleic Acids Res.* **33**(17): p. 5691-702.

84 ^{GTL}

Metabolic Engineering of Light and Dark Biochemical Pathways in Wild-Type and Mutant *Synechocystis* PCC 6803 Strains for Maximal, 24-Hour Production of Hydrogen Gas

P. S. Schrader,¹ E. H. Burrows,² F. W. R. Chaplen,² and R. L. Ely^{2*} (ely@engr.orst.edu)

¹Yale University, New Haven, Connecticut and ²Oregon State University, Corvallis, Oregon

Project Goals: The objectives of the proposed research are addressed in the following four tasks:

1. Evaluate the effects of various culture conditions (N, S, or P limitation; light/dark; pH; exogenous organic carbon) on H₂ production profiles of WT cells and an NDH-1 mutant; 2. Conduct metabolic flux analyses for enhanced H₂ production profiles using selected culture conditions and inhibitors of specific pathways in WT cells and an NDH-1 mutant; 3. Create PCC 6803 mutant strains with modified H₂ases exhibiting increased O₂ tolerance and greater H₂ production; 4. Integrate enhanced H₂ase mutants and culture and metabolic factor studies to maximize 24-hour H₂ production.

Global power demand is projected to increase from the current 13 terawatts (TW) to 46 TW by the end of this century (US DOE, 2005). Though fossil fuels will certainly continue to dominate the growth of the energy sector in the near term (E.I.A., 2006), a renewable alternative must be pursued now to prevent severe economic disruptions from a transition that may take decades to complete (Hirsch, 2005). Solar energy, harvested directly in the form of hydrogen gas (H₂) from the splitting of water, offers a promising alternative to meet regional, national, and global energy demand in a sustainable, environmentally friendly manner.

We are using the cyanobacterial species *Synechocystis* sp. PCC 6803 to pursue two initial project goals:

1) Optimize H₂ production conditions through a series of H₂ profiling experiments varying the light/

dark ratio, pH, exogenous organic carbon, and nitrogen, sulfur, and phosphorus concentrations, and 2) Conduct metabolic flux analysis to quantify intracellular reductant fluxes and to identify obstructions to H_2 production, specifically to identify factors that decrease intracellular levels of NADPH to support H_2 production. Both of these goals are being addressed with wild-type cells as well as with a high H_2 -producing mutant with impaired type I NADPH-dehydrogenase (NDH-1) function.

Results from the **H_2 profiling experiments** show that pH appears to have a dramatic effect on H_2 production by *Synechocystis* sp. PCC 6803. Over pH values ranging from 5 to 10, we have observed significantly higher H_2 production at higher pH, both by WT cells and by the NDH-1 mutant. With WT cells, this effect seems especially pronounced when N is high (18 mM or 35 mM), whereas high P (180 μ M and 360 μ M) seems to dampen this effect. In contrast, with the NDH-1 mutant, the effect is pronounced when N is limiting, and we have seen mixed results with P. At higher pH, there may be a trend toward higher H_2 production at higher P concentrations, but the data we have generated so far have not shown it to be statistically significant. At neutral pH, our preliminary results suggest that slightly higher H_2 production may be occurring under N-limited conditions in WT cells, which would be consistent with the observations of Schutz et al. (2004), who found that nitrogen-limitation (1mM N) increased H_2 photoproduction in *Synechocystis* sp. PCC 6803 over 48 hours. Results of screening tests conducted so far have not shown a similar trend in the NDH-1 mutant. Also, any tendencies toward higher H_2 production caused by N limitation seem to be overwhelmed by pH effects at higher pH values, as described above. Our initial results suggest that the optimal concentration of $NaHCO_3$ for H_2 production, from the concentrations tested, is 80 mM. Tsygankov et al. (2002) found that sulfur deprivation increased H_2 photoproduction in *Chlamydomonas reinhardtii*, so we have anticipated that sulfur deprivation may have a similar effect in *Synechocystis* PCC 6803. Preliminary results regarding the effects of S concentration on H_2 production have been mixed, and we are looking into this question further. In evaluating light/dark effects, we have tested 12 and 24 light-dark cycles per day so far, with the dark/light ratio varying from 8:1 to 29:1 based on published values of the respiration/photosynthesis ratio. Initial results suggest that a regimen consisting of 24 cycles per day with a dark/light ratio of 8:1 produces the highest H_2 production. We are working on an advanced optimization algorithm that will predict the optimal conditions for hydrogen production, based on the data collected.

Metabolic flux analysis is a linear algebra technique used to determine the fluxes among a network of intracellular metabolites, based on measured inflows and outflows from the cells, and based on the assumption that the system is at steady-state. To achieve true metabolic steady-state, we are conducting the metabolic flux analysis experiments in a chemostat. We are working to achieve a sinusoidal steady-state because the chemostat is exposed to light/dark cycling, which is needed for H_2 production. The network chosen for photoautotrophic growth consists of 20 constraints and 24 reactions, thus requiring at least 4 measurements. The inflows and outflows measured are H_2 , O_2 , CO_2 , glucose, glycogen, ammonium, and biomass production/consumption. Intracellular NADPH/NADP⁺ concentrations and light intensity are also measured.

References

1. E.I.A. 2006. International Energy Outlook 2006, Energy Information Administration, U.S. Department of Energy.
2. Hirsch, R. L. 2005. Testimony on Peak Oil, House Subcommittee on Energy and Air Quality, U.S. House of Representatives.
3. Schutz, K., T. Happe, O. Troshina, P. Lindblad, E. Leita, P. Oliveira, & P. Tamagnini. 2004. Cyanobacterial H_2 production - a comparative analysis. *Planta*. **218**(3):350-359

4. Tsygankov, A., S. Kosourov, M. Seibert, & M.L. Ghirardi. 2002. Hydrogen photoproduction under continuous illumination by sulfur-deprived, synchronous *Chlamydomonas reinhardtii* cultures. *Int. J. Hydrogen Energy*. **27**(11-12):1239-1244
5. U.S. DOE. 2005. Basic Research Needs for Solar Energy Utilization. Report of the Basic Energy Sciences Workshop on Solar Energy Utilization. US Department of Energy. Argonne National Laboratory Publication 2005.

85 [—]GTL

Pathway Tools + MetaCyc = Comprehensive Pathway Modeling

Ron Caspi,¹ Hartmut Foerster,² Carol Fulcher,¹ Michelle Green,¹ Pallavi Kaipa,¹ Markus Krummenacker,¹ Mario Latendresse,¹ Suzanne Paley,¹ Chris Tissier,² Peifen Zhang,² Sue Rhee,² and **Peter D. Karp**^{1*} (pkarp@ai.sri.com)

¹Bioinformatics Research Group, SRI International, Menlo Park, California and ²Carnegie Institution, Department of Plant Biology, Stanford, California

Project Goals: The goal of the MetaCyc project is to develop a comprehensive, universal database of experimentally derived information on metabolic pathways and enzymes from many organisms. The goal of the Pathway Tools project is to facilitate understanding of metabolic and regulatory networks, and their relationship to the genome, by developing computational tools for inference, querying, visualization, and analysis of databases that integrate pathway and genome information.

Metabolic engineering demands an accurate model of the metabolic network of a target organism and the relationship of that network to the genome, plus powerful analysis tools for constructing, refining, and analyzing that model.

The MetaCyc multiorganism pathway database [1,2] describes experimentally elucidated metabolic pathways and enzymes reported in the experimental literature. MetaCyc is both an online reference source on metabolic pathways and enzymes for metabolic design, and a solid foundation of experimentally proven pathways for use in computational pathway prediction. MetaCyc version 10.6 describes 890 pathways from more than 900 organisms. The 6100 biochemical reactions in MetaCyc reference 6000 chemical substrates, most of which contain chemical structure information. MetaCyc describes the properties of 3500 enzymes, such as their subunit structure, cofactors, activators, inhibitors, and in some cases their kinetic parameters. The information in MetaCyc was obtained from 12,000 research articles, and emphasizes pathways and enzymes from microbes and plants, although it also contains animal pathways.

Pathway Tools [3,4] constructs a metabolic model of an organism from its annotated genome using the following computational inference tools. The model is in the form of a Pathway/Genome Database (PGDB).

- It predicts the metabolic pathways of the organism by recognizing known pathways from the MetaCyc database
- It predicts which genes fill holes in those metabolic pathways (pathway holes are pathway steps for which no enzyme has been identified in the genome)
- It predicts operons for prokaryotic genomes

- It infers the presence of transport reactions from the names of transport proteins in the genome annotation
- The software automatically generates a one-screen cellular overview diagram containing the metabolic and transport networks of the cell

A set of graphical editors within Pathway Tools allows scientists to refine a PGDB by adding, or modifying metabolic pathways, gene annotations, reactions, substrates, and regulatory information. The existence of an accurate knowledge base of the metabolic network is a critical resource for metabolic engineering.

The software provides a large number of operations for querying, visualization, web publishing, and analysis of PGDBs. New network debugging tools allow the user to find errors or incompleteness in the metabolic model by identifying dead-end metabolites, and mismatches between the transport and metabolic subsystems. These tools can speed the identification of errors in the genome annotation and in the metabolic model.

A new metabolite tracing tool supports graphical exploration of the path that a substrate follows through the metabolic network, in either the forward or backward direction. The user interactively guides the software in selecting which branches of metabolism to follow, and metabolic paths are highlighted on the cellular overview diagram. A new tool for graphical construction of complex database queries provides a quantum leap in the power of database queries that a user can construct without knowledge of SQL.

A family of omics viewers support systems-level visualization of large-scale datasets onto cellular networks. The first omics viewer paints omics datasets onto the cellular overview of metabolic and transport networks. The second (new) omics viewer paints omics datasets onto a diagram of the transcriptional regulatory network. The third (new) omics viewer paints omics datasets onto the genome. These tools provide complementary perspectives for interpreting omics data.

A set of new comparative genomics tools supports many comparisons across the genomes and metabolic networks of a set of organism's PGDBs. For example, the pathway complements of selected PGDBs can be compared, with the results ordered by pathway ontology.

Other visualization tools include automated display of metabolic pathways, reactions, enzymes, genes, and operons, and a genome browser.

More than 75 groups are using Pathway Tools and MetaCyc to produce PGDBs for more than 150 organisms, including the major model organisms for biomedical research (yeast, worm, fly, *Dictyostelium*), pathogens of biodefense interest, GTL organisms, many other bacteria and archaea, and plants (including *Arabidopsis*, *Medicago*, Rice, Tomato, and Potato).

References

1. R. Caspi et al, "MetaCyc: A multiorganism database of metabolic pathways and enzymes," *Nucleic Acids Research* 34:D511-6 2006.
2. P. Zhang et al, "MetaCyc and AraCyc. Metabolic Pathway Databases for Plant Research," *Plant Physiol* 138:27-37 2005.
3. P.D. Karp et al, "The Pathway Tools Software," *Bioinformatics* 18:S225-32 2002.
4. Paley, S.M. et al, "The Pathway Tools Cellular Overview Diagram and Omics Viewer," *Nucleic Acids Research* 34:3771-8, 2006.
5. Paley, S.M. et al, "Creating Fungal Pathway/Genome Databases Using Pathway Tools," *Applied Mycology and Biotechnology* 6:209-26 2006.

Constraint-Based Modeling of Central Metabolism in the Family *Geobacteraceae*

Jun Sun^{1*} (jsun@genomatica.com), Steve Van Dien,¹ Radhakrishnan Mahadevan,² Maddalena Coppi,³ Laurie DiDonato,³ Carla Risso,³ Mounir Izallalen,³ Bradley Postier,³ Raymond DiDonato,³ Kai Zhuang,² Priti Pharkya,¹ Tom Fahland,¹ Olivia Bui,¹ Iman Famili,¹ Christophe Schilling,¹ and **Derek Lovley³**

¹Genomatica Inc., San Diego, California; ²University of Toronto, Toronto, Ontario; and ³University of Massachusetts, Amherst, Massachusetts

Project Goals: The ultimate goal of the *Geobacter* Project is to develop genome-based in silico models that can be used both to interpret environmental gene expression data in environments in which *Geobacteraceae* predominate and to predict the growth and metabolism of *Geobacteraceae* in situ using routine geochemical measurements as input. Such models will also enable prediction of the outcome of various potential manipulations that might be made to optimize processes of interest, such as in situ uranium bioremediation and harvesting electricity from waste organic matter, prior to conducting costly and labor-intensive field experiments.

As part of an effort to predictively model the behavior of *Geobacteraceae* involved in bioremediation and electricity-generation *in situ*, *in silico* reconstructions of the metabolic networks of three *Geobacter* species were generated by the constraints-based approach.

The genome-scale metabolic model of *Geobacter sulfurreducens*, the genetically tractable and acetate-oxidizing species, was the first to be generated. In the past year, ¹³C isotopic labeling studies were performed to quantitate actual flux distributions and to further refine the *G. sulfurreducens* model. Based on the labeling patterns from acetate:fumarate chemostat cultures, overall amino acid labeling patterns were consistent with flux distributions generated by the model. Computational flux analysis on the levels of flux through the various phosphoenol pyruvate (PEP) synthesizing pathways indicated that PEP was derived from both acetate and oxaloacetate, despite the energetic differences. One surprising prediction by modeling was an ATP-consuming futile cycle that involved three enzymes catalyzing the interconversion of pyruvate, oxaloacetate and PEP. The futile cycle was confirmed by the labeling data. This futile cycle may only be active during growth on fumarate, which is not a natural electron acceptor. The possibility that adaptive evolution on fumarate will lead to the inactivation of this futile cycle is currently under investigation.

¹³C-labeling studies also led to the discovery of an alternate pathway for the biosynthesis of isoleucine in *G. sulfurreducens*. The metabolic model predicted that isoleucine was synthesized from aspartate and pyruvate, but the labeling pattern of isoleucine did not match the prediction. Further flux analysis suggested that the majority of isoleucine was synthesized exclusively from pyruvate and/or acetyl-CoA potentially via a citramalate pathway that was found in several species of methanogens and *Leptospira interrogans*. The citramalate synthase activity was indeed detected in soluble *G. sulfurreducens* extracts and a candidate citramalate synthase gene was identified. Optimizations performed to assess the relative contributions of the citramalate and the aspartate/pyruvate pathways to isoleucine biosynthesis indicated that 68-78% of the isoleucine was synthesized via the citramalate pathway. In addition, simulations indicated that the use of the citramalate pathway instead of the aspartate/pyruvate pathway significantly increased the efficiency of isoleucine biosynthesis during growth on acetate.

As part of an ongoing investigation of the effects of nutrient limitation and other environmentally relevant stresses on central metabolism, the metabolic model was used to estimate the energetic cost of nitrogen fixation and to predict changes in flux distribution as a result of nitrogen fixation. Studies comparing the effects of nitrogen fixation on global gene expression to predicted changes in flux distribution are underway, and a similar approach is being applied to phosphate limitation. The effect of growth rate on central metabolism is also being investigated, as the growth rates of *Geobacteraceae in situ* are much lower than those typically observed in laboratory cultures. *G. sulfurreducens* was cultivated in chemostats at a variety of growth rates spanning both the low and high end of the spectrum and changes in gene expression in response to changes in growth rate are being compared to predicted changes in flux distribution.

The engineering of *Geobacter* species to achieve increased respiration rates as a strategy for increasing electricity production is another area of investigation. Respiration rates were elevated in *G. sulfurreducens* by inducing expression of an ATP-consuming enzyme, the peripheral subunits of the F1F0-ATPase. *In silico* analysis of the metabolic network, using experimentally derived organic acid measurements as input, indicated that introduction of this ATP-drain should result in diversion of acetate from biosynthetic reactions to the TCA cycle and ATP generation. In order to assess the accuracy of these *in silico* predictions, microarray analysis was performed following induction of the ATP drain. There were many examples of changes in gene expression which were consistent with *in silico* predictions. A variety of genes involved in energy metabolism, including multiple cytochromes and electron transport proteins, TCA cycle enzymes, and subunits of the NADH dehydrogenase, were up-regulated, as was the gene encoding the dicarboxylic acid exchanger involved in fumarate uptake. In contrast, the gene encoding acetate kinase, which activates acetate for gluconeogenesis was down-regulated.

Considerable progress has been made towards the development and refinement of metabolic models for two other *Geobacteraceae*: *G. metallireducens* and *P. carbinolicus*, both of which have metabolic capabilities that are not shared by *G. sulfurreducens*. The *G. metallireducens* and *P. carbinolicus* models were initially created using an automated model reconstruction procedure, the Automodel pipelineTM and were manually curated over the past year. It was estimated that the use of the Automodel pipelineTM accelerated model development by 3.7 fold. The Automodel pipeline was also used to rapidly incorporate discoveries made during development of the two newer models into the *G. sulfurreducens* model, resulting in an increase in the total number of reactions and a decrease in the number of non-gene associated reactions. Currently, the *G. sulfurreducens*, *G. metallireducens* and *P. carbinolicus* models contain 649, 606 and 700 reactions, respectively. More than half (64% to 74%) of the reactions present in each individual model are shared among all three models. The unique reactions in the *P. carbinolicus* model include those involved in fermentation and proline biosynthesis, whereas unique reactions in the *G. metallireducens* model include those involved in the metabolism of monoaromatic compounds. In addition, the *G. metallireducens* and *P. carbinolicus* models contain a key alcohol dehydrogenase for ethanol utilization and the oxidative branch of the pentose phosphate pathway, which are not present in the *G. sulfurreducens* model. *In silico* deletion analysis was performed on all three models and revealed that ca. 200 reactions were essential for fermentative growth of *P. carbinolicus* on acetoin and respiratory growth of either *G. metallireducens* and *G. sulfurreducens* on acetate and Fe(III) citrate. Further computational and experimental analyses using these models will provide insight into the metabolism of these and other species of *Geobacteraceae*.

Finally, as microorganisms rarely exist in isolation in the environment, computational analysis of microbial communities has also been initiated. Microbial communities in which Fe(III) serves as the terminal electron acceptor typically include fermentative organisms, which break down complex organic matter, and *Geobacteraceae* which utilize fermentation byproducts as electron donors for

Fe(III) reduction. In order to model a simplified version of such a community, a coculture consisting of *Escherichia coli* and *G. sulfurreducens*, which together can couple glucose fermentation to iron reduction, has been established. A dynamic model of this co-culture has been developed, and predictions generated by this model have been reconciled with physiological data.

87 [—]GTL

Analysis of Degree of Genetic Redundancy in Prokaryotic Metabolic Networks

R. Mahadevan^{1*} (mahadevan@chem-eng.utoronto.ca) and D.R. Lovley²

¹University of Toronto, Ontario and ²University of Massachusetts, Amherst, Massachusetts

Project Goals: Metabolic networks can be robust to environmental and genetic perturbations through genetic redundancy or alternate biochemical pathways. The role of these factors has been analyzed extensively in eukaryotes such as *Saccharomyces cerevisiae* and recent studies indicate a stronger role for gene duplicates in accentuating enzymatic flux rather than as a back-up function. Here, we analyze the extent of genetic and biochemical redundancy in prokaryotic metabolic networks using genome-scale metabolic models. Surprisingly, we find that the extent of genetic redundancy appears to be enriched in *Geobacter sulfurreducens* and *Methanosarcina barkeri* as compared to other organisms. Based on these findings, we suggest that the environmental niche, an organism inhabits might have a role in determining the mechanism of attaining robustness to genetic perturbations.

Robustness to perturbations is almost an intrinsic and essential component of biological systems. Several studies have extensively analyzed the robustness of biological systems and have provided insights on the mechanism of adaptation to genetic and environmental perturbations. Robustness of metabolism after the loss of a gene product can occur due to either the presence of a gene duplicate that has the same function (“gene family buffering”) or due to the presence of alternate pathways that can achieve similar function (“pathway buffering”). The roles of genetic redundancy and biochemical buffering and the mechanistic principles arising from molecular interactions is of great interest to further understand robustness in biological systems. Such an understanding of the factors favoring the maintenance of duplicate genes in microbial genomes is essential for developing models of microbial evolution.

A genome-scale flux-balance analysis of the metabolic network of *Saccharomyces cerevisiae* has suggested that gene duplications primarily provide increased enzyme dosage to enhance metabolic flux because the incidence of gene duplications in essential genes is no higher than that in non-essential genes. However, *S. cerevisiae* represents just one example of a wide spectrum of microbial metabolism and is a eukaryote. Therefore, we used genome-scale metabolic models of *Escherichia coli*, *Bacillus subtilis*, *Geobacter sulfurreducens* and *Methanosarcina barkeri* to analyze the extent of genetic and biochemical redundancy in prokaryotes that are either specialists, with one major mode of energy generation, or generalists, which have multiple metabolic strategies for conservation of energy.

Genome-scale metabolic models represent the majority of the biological information ranging from genome sequence, biochemical and high-throughput physiological data and have been shown to be successful in predicting the experimentally determined deletion phenotypes. Surprisingly, the results suggest that although generalists, such as *E. coli* and *B. subtilis*, are similar to the eukaryotic general-

ist, *S. cerevisiae*, in having a low percentage (< 10 %) of essential genes and few duplications of these essential genes, metabolic specialists, such as *G. sulfurreducens* and *M. barkeri*, have a high percentage (> 30 %) of essential genes and a high degree of genetic redundancy in these genes compared to non-essential genes. The analysis of flux through the reactions with the gene duplicates reveal that they are no more likely to have a higher rate of flux than the rest of reactions further suggesting a different role for gene duplicate in specialists such as *G. sulfurreducens*. Therefore, the specialist organisms appear to rely more on gene duplications rather than alternative-but-equivalent metabolic pathways to provide resilience to gene loss. Generalists rely more on alternative pathways. Thus, the concept that the role of gene duplications is to boost enzymatic flux rather than provide metabolic resilience, may not be universal. Rather, the degree of gene duplication in microorganisms may be linked to mode of metabolism and environmental niche.

88 ^{GTL}

Mechanisms of Sulfur Reduction by *Shewanella*

Edward J. Crane III^{1*} (EJ.Crane@pomona.edu), Evan T. Hall,¹ and Ken Nealson²

¹Pomona College, Claremont, California and ²University of Southern California, Los Angeles, California

Project Goals: This project is a component of the *Shewanella* Federation and as such contributes to the overall goal of applying the tools of genomics, leveraging the availability of genome sequence for 18 additional strains of *Shewanella*, to better understand the ecophysiology and speciation of respiratory-versatile members of this important genus. To understand these systems the SF is using genome-based approaches to investigate *Shewanella* as a system of integrated networks; first describing key cellular subsystems those involved in signal transduction, regulation, and metabolism—then building towards understanding the function of whole cells and, eventually, cells within populations. As a general approach, the SF is collectively employing complimentary top-down bioinformatics-based genome functional predictions, high-throughput expression analyses, and functional genomics approaches to uncover key genes as well as metabolic and regulatory networks. The bottom-up component employs more traditional approaches including genetics, physiology and biochemistry to test or verify predictions. This information will ultimately be linked to analyses of signal transduction and transcriptional regulatory systems and used to develop a linked model that will contribute to understanding the ecophysiology of *Shewanella* in redox stratified environments.

Shewanella is famed for its ability to respire a wide range of substrates, and sulfur is one of the substrates that many species of this genus are able to use as an electron acceptor. We are interested in characterizing the mechanisms by which *S. oneidensis* MR-1 reduces sulfur, and are characterizing the sulfur reductase complex of this organism. We will determine growth yields and electron budgets for lactate oxidized/S⁰ reduced (with an N₂ atmosphere, as we have shown that H₂ is not required for growth on S⁰), lactate/S⁰ with H₂ (as an additional electron donor), and with electron donors such as N-acetylglucosamine, α-ketobutyrate and α-ketoglutarate. We will also characterize the mechanism of the putative polysulfide reductase complex, using a range of spectroscopic techniques. One of the difficulties associated with the *in vitro* growth of *Shewanella* with sulfur as an electron acceptor is the inhibition of the organism by the sulfide product. Simply trapping sulfide is not an option, as sulfide is also necessary to begin the reduction process by reducing insoluble sulfur to soluble polysulfide (which is believed to be the actual substrate for the sulfur reductase). We have determined

that optimal sulfide concentrations for anaerobic growth are in the range of 1 mM sulfide. We have isolated membranes from MR-1 growing with sulfur, and are in the process of developing a sulfur reductase assay, as well as a method for suspending the sulfur reductase complex in polyanionic films in order to facilitate spectroscopic characterization. While comparison of the genome of several *Shewanella* species to those of other sulfur reducers has revealed the enzymes most likely to be involved in the direct reduction of S^0 and/or polysulfide, it is much less clear which pathways feed electrons to these enzymes and how the enzymes and complexes within the system interact. Extensive screening of mutants of the c-type cytochromes has revealed the majority of these proteins, including the Mtr and Omc heme proteins involved in metal reduction, as well as $\Delta SO4144$, which has been shown to be essential for the reduction of tetrathionate, are not essential for sulfur reduction. Of these proteins, only $\Delta SO2930$ and $\Delta SO2931$ appear to show any decrease in ability to reduce sulfur.

89 ^{GTL}

Carbon and Energy Metabolism Strategies in *Shewanella*

G. Pinchuk^{1*} (Grigoriy.Pinchuk@pnl.gov), A. Beliaev,¹ O. Geydebrekht,¹ D. Kennedy,¹ I. Famili,² J. Reed,² J. Scott,³ S. Reed,¹ M. Romine,¹ and J. Fredrickson¹

¹Pacific Northwest National Laboratory, Richland, Washington; ²Genomatica, Inc., San Diego, California; and ³Earth Sciences, Dartmouth College, Hanover, New Hampshire

Project Goals: This project is a component of the *Shewanella* Federation and as such contributes to the overall goal of applying the tools of genomics, leveraging the availability of genome sequence for 18 additional strains of *Shewanella*, to better understand the ecophysiology and speciation of respiratory-versatile members of this important genus. To understand these systems the SF is using genome-based approaches to investigate *Shewanella* as a system of integrated networks; first describing key cellular subsystems those involved in signal transduction, regulation, and metabolism—then building towards understanding the function of whole cells and, eventually, cells within populations. As a general approach, the SF is collectively employing complimentary top-down bioinformatics-based genome functional predictions, high-throughput expression analyses, and functional genomics approaches to uncover key genes as well as metabolic and regulatory networks. The bottom-up component employs more traditional approaches including genetics, physiology and biochemistry to test or verify predictions. This information will ultimately be linked to analyses of signal transduction and transcriptional regulatory systems and used to develop a linked model that will contribute to understanding the ecophysiology of *Shewanella* in redox stratified environments.

The genus *Shewanella* is unusually well-adapted to chemically (redox) stratified environments as reflected in the ability to utilize a broad range of electron acceptors via a highly diversified electron transport system. Occupying such niches requires the ability to adapt rapidly to changes in electron donor/acceptor type and availability; hence the ability to compete and thrive in such environments must ultimately be reflected in the organization and flexibility of the electron transfer networks as well as central carbon metabolism pathways. Although MR-1 is typically considered to have a relatively restricted substrate range for carbon and energy sources, genome-based analyses revealed multiple pathways for C_{2-6} compounds, amino acids, and fatty acids, reflecting its ecological role as a consumer of organic matter breakdown products in relatively carbon-rich environments that support diverse anaerobic microbial communities. Using controlled cultivation, biochemical, genetic and

genomic approaches in conjunction with pathway modeling, we showed that (i) metabolic pathways in MR-1 expressed under different redox conditions utilize pyruvate as a key metabolite and (ii) the pathways involved in ATP production under aerobic and anaerobic conditions fundamentally differ reflecting the amount of energy this organism can generate by oxidative phosphorylation.

Aerobic utilization of C₂₋₅ compounds through pyruvate as a central intermediate. When grown on lactate or pyruvate (which is the first product of lactate oxidation) *S. oneidensis* MR-1 displayed the highest growth rates in comparison to other substrates tested. To determine which metabolic pathways are used to metabolize different organic compounds, several MR-1 deletion mutants with genes predicted to be involved in central carbon metabolism were generated. Initially, a pyruvate dehydrogenase complex (PDHc) mutant carrying a deletion of the E1 subunit (SO0424) was tested for aerobic growth. We determined that this mutant was unable to grow on minimal medium supplemented with any single C₂₋₅ compounds tested including acetate. Results obtained with other mutants tested for their ability to grow with lactate or acetate revealed that isocitrate lyase (SO1484) and malate synthase (SO1483) are not essential for growth with lactate but are indispensable for acetate metabolism. Malic enzyme (SO3855) was not required for aerobic growth with any compound tested. Acetyl-CoA synthase, but not the combined action of phosphotransacetylase and acetate kinase, was necessary for exogenous acetate utilization. Taken together, these results suggest that pyruvate is the central metabolic intermediate involved in aerobic utilization of C₂₋₅ compounds by *S. oneidensis* MR-1. Additionally, our experiments strongly suggest that *S. oneidensis* does not oxidize acetate using the TCA cycle under aerobic conditions, and a new pathway responsible for aerobic acetate oxidation in *S. oneidensis* cells is proposed.

The role of substrate-level phosphorylation under anaerobic and O₂-limited growth of *S. oneidensis* MR-1.

One of the fundamental characteristics of *S. oneidensis* metabolism is its inability to use acetate as an electron donor under anaerobic conditions. We extended these previous observations by demonstrating that acetate cannot be used as carbon and energy source by *S. oneidensis* MR-1 under Fe(III)- and fumarate-reducing as well as under O₂-limited conditions. Anaerobic or O₂-limited growth with lactate as electron donor is accompanied by acetate excretion (with 80-90% lactate converted to acetate). These results implied that acetate excretion may be coupled to ATP production catalyzed by acetate kinase (SO2915). Indeed, an MR-1 acetate kinase deletion mutant did not grow anaerobically with either Fe(III)-NTA or fumarate when lactate served as the carbon and energy source. Chemostat experiments also showed that the amount of acetate produced was in inverse proportion to the O₂ flux. These results strongly suggest that under anaerobic or O₂-limited growth *S. oneidensis* MR-1 depends solely on substrate level phosphorylation for energy generation.

We also have generated several lines of evidence, including analysis using a flux balance model of *S. oneidensis* metabolism, which show that under conditions of O₂ limitation and fumarate reduction most of ATP is produced from lactate or pyruvate on the level of substrate phosphorylation (from acetyl phosphate). In contrast, the redox chain functions mostly to re-oxidize electron carriers and, in case of fumarate, does not couple electron acceptor reduction to oxidative phosphorylation. Our results suggest that for *S. oneidensis* the rate of electron transfer to a terminal electron acceptor determines the growth rate and part of energy spent on maintenance needs, whereas efficiency of electron transport coupling to phosphorylation partially determine biomass growth yield. Such flexibility of central carbon metabolism allows *Shewanella* to survive during periods of nutrient-limitation and proliferate rapidly when both electron acceptor(s) and donor(s) are available.

Metabolic Reconstruction of *Shewanella oneidensis*: A Community Resource

Jennifer L. Reed,¹ Iman Famili,² **Sharon J. Wiback**,² Christophe H. Schilling,² Grigoriy Pinchuk,³ Margaret R. Romine,³ Johannes C. Scholten^{3*} (johannes.scholten@pnl.gov), Joel Klappenbach,⁴ and James K. Fredrickson³

¹Department of Bioengineering, University of California, San Diego, La Jolla, California;

²Genomatica, Inc., San Diego, California; ³Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington; and ⁴Center for Microbial Ecology, Michigan State University, East Lansing, Michigan

Project Goals: This project is a component of the *Shewanella* Federation and as such contributes to the overall goal of applying the tools of genomics, leveraging the availability of genome sequence for 18 additional strains of *Shewanella*, to better understand the ecophysiology and speciation of respiratory-versatile members of this important genus. To understand these systems the SF is using genome-based approaches to investigate *Shewanella* as a system of integrated networks; first describing key cellular subsystems - those involved in signal transduction, regulation, and metabolism—then building towards understanding the function of whole cells and, eventually, cells within populations. As a general approach, the SF is collectively employing complementary “top-down” - bioinformatics-based genome functional predictions, high-throughput expression analyses, and functional genomics approaches to uncover key genes as well as metabolic and regulatory networks. The “bottom-up” component employs more traditional approaches including genetics, physiology and biochemistry to test or verify predictions. This information will ultimately be linked to analyses of signal transduction and transcriptional regulatory systems and used to develop a linked model that will contribute to understanding the ecophysiology of *Shewanella* in redox stratified environments.

Genome-scale network reconstructions account for components and component interactions in biological networks, and are a way in which to collate and analyze data from a variety of sources. Here we report a metabolic reconstruction for *Shewanella oneidensis* MR-1 based on the current genome annotation and primary literature. The reconstruction includes 769 reactions, 779 genes, and 660 metabolites. The reconstruction was used to build a flux balance model that was used in a variety of computational analyses, including: assessment of growth phenotypes, evaluation of metabolite usage (as substrates or by-products), and prediction of knock-out phenotypes. The model correctly predicted growth on a variety of carbon and nitrogen sources. In addition, quantitative evaluation of alternative electron acceptors led to the identification of 7 classes of electron acceptors, with differing biomass yields (g D.W. produced per mmol electron acceptor consumed). Gene deletion simulations across 10 different environmental conditions with various carbon sources and electron acceptors found that a large fraction of genes were never essential (535 out of 779), while a smaller fraction were always essential (202 out of 779) for growth on these 10 conditions. Together this work provides a resource that can be used by *Shewanella* researchers and illustrates how reconstructions can serve as a means to evaluate experimental data and generate testable hypotheses to better understand its ecophysiology.

91 ^{GTL}**The Challenge of Incorporating Regulatory Effect in Genome-Scale Networks**

C.L. Barrett*, M.J. Herrgard*, B.K. Cho, E.M. Knight, J. Elkins, and **B.O. Palsson** (palsson@ucsd.edu)

Department of Bioengineering, University of California, San Diego, California

Project Goals: Genomic and bibliomic data has been used to reconstruct a number of genome-scale metabolic networks. The stoichiometric structure of these networks has enabled a series of basic and applied studies that address both proximal and distal causation in biology. One of the challenges going forward with computational models at the genome-scale is to account for regulatory effects. Regulation occurs primarily at two levels; 1) the transcriptional level, and 2) the gene product activity level. Significant progress is being made with the former issue, while the latter is still at a conceptual stage. The current state and future challenges of both issues will be discussed.

Genomic and bibliomic data has been used to reconstruct a number of genome-scale metabolic networks. The stoichiometric structure of these networks has enabled a series of basic and applied studies that address both proximal and distal causation in biology. One of the challenges going forward with computational models at the genome-scale is to account for regulatory effects. Regulation of metabolic enzymes occurs primarily at two levels; 1) the transcriptional level, and 2) the post-transcriptional (protein expression and activity) level. Significant progress is being made with characterizing, reconstructing and modeling transcriptional regulatory networks regulating metabolism. However, the development of the corresponding methods for incorporating post-transcriptional regulation into genome-scale models is still at an early stage. The current state and future challenges of incorporating both transcriptional and post-transcriptional regulation in genome-scale networks will be discussed.

92 ^{GTL}**Acclimation of *Chlamydomonas reinhardtii* to Anoxic Conditions: Gene Expression, Hydrogenase Induction and Metabolic Pathways**

Michael Seibert^{1*} (mike_seibert@nrel.gov), Florence Mus,² Alexandra Dubini,^{1,3} Maria L. Ghirardi,¹ Matthew C. Posewitz,³ and Arthur R. Grossman²

¹National Renewable Energy Laboratory, Golden, Colorado; ²Carnegie Institution of Washington, Stanford, California; and ³Colorado School of Mines, Golden, Colorado

Project Goals: Past research has shown that photosynthesis, respiration, and fermentation are all required to sustain H₂ photoproduction from water in algae. These microbes utilize [Fe]-hydrogenases, which are the most efficient H₂-generating biocatalysts known. The long-term objective of our project is to identify the suite of genes facilitating and/or limiting H₂ photoproduction in the alga, *Chlamydomonas reinhardtii*, by conducting global gene expression and cell metabolism studies using algal cells acclimated to conditions known to induce H₂-production activity. A detailed understanding of the influences of metabolism and other environmental factors on

the coordinated expression of genes and biochemical pathways associated with H₂-production activity will ultimately be required to increase the yields of renewable H₂ production for potential future applications. To accomplish this we will examine WT cells and a number of NRELs H₂-production mutants under a number of experimental conditions using *Chlamydomonas* gene microarrays along with extensive biochemical assays. Algal H₂ production requires the synergies of multiple redox proteins, sensors, biochemical pathways and regulatory processes. Knowledge gained by deconvoluting these interactions will help us identify specific targets for future strain engineering aimed at enhancing H₂ production in *C. reinhardtii*.

The unicellular green alga *Chlamydomonas reinhardtii* has emerged as a prototype organism for investigating processes such as photosynthesis, nutrient deprivation, flagellar function, and H₂ production. Although previous physiological studies have linked fermentation and photosynthetic electron transport to H₂ production in *C. reinhardtii*, a more precise knowledge of the metabolic and regulatory context required for H₂ production will be necessary in order to understand current limitations in H₂ yields. We have combined molecular and physiological approaches to examine the acclimation of *C. reinhardtii* strain CC-425 during a shift from oxic to anoxic conditions, which leads to H₂ evolution. The levels of transcripts involved in fermentative metabolism were monitored to determine whether the accumulation of these transcripts reflects the abundance of specific metabolites that accumulate in the cultures during anoxic adaptation. We also used high-density, oligonucleotide-based microarrays to obtain insights into the genome-wide responses initiated by anoxia, as monitored by changes in the relative abundance of ~10,000 unique transcripts. While transcripts from a number of genes associated with fermentation metabolism increased, as expected, several genes encoding proteins involved in transcriptional/translational regulation, post-translational modifications, and stress responses also increased as the cell cultures became anoxic.

Microarray and qPCR analyses were used to examine the pathways associated with fermentation at the RNA level. Several transcripts encoding proteins critical for fermentation increase as the cultures became anoxic. Indeed, anoxia leads to the upregulation of transcripts encoding pyruvate formate lyase (PFL), pyruvate:ferredoxin oxidoreductase (PFOR), alcohol dehydrogenase (ADHE), phosphotransacetylase (PTA2) and acetate kinase (ACK1), as well as some cognate proteins. These results imply that upon exposing cells to dark, anoxic conditions, *C. reinhardtii* can switch very rapidly to fermentative metabolism. The fermentation products synthesized by *C. reinhardtii* following the imposition of anoxia include malate, formate, acetate, and ethanol. Formate, acetate and ethanol, in the ratio 1:1:0.5, were the major fermentative products formed over a 24-h period of anoxia; malate was observed only at minor levels. The observed ratio of fermentation products confirms that both the PFL and PFOR pathways are activated as O₂ in the cultures declines. Moreover, during fermentation, starch, the principal carbon-storage compound in this alga, is degraded primarily to glucose-1-phosphate, which is subsequently oxidized to pyruvate during glycolysis. Accordingly, the levels of transcripts encoding amylase and β -amylase (involved in starch degradation) also increased.

The array data hints at regulatory processes that accompany the acclimation of the cells to dark, anoxic conditions. Many transcripts that increase correspond to chloroplast regulatory elements, including ppGpp synthetase/degradase, Mbb1 factor, translation initiation factor IF2, Tab2, and Tbc2 proteins. Elevated levels of transcript for these polypeptides may indicate the need to control translational and post-translational processes that occur in the chloroplast as the environment becomes anoxic (perhaps reflecting both structural and functional changes that occur in the chloroplast). Transcripts, encoding several proteases and kinases, also increase in cells following exposure to dark, anoxic conditions. These results suggest the possible activation of specific signalling pathways, the initiation of specific protease-dependent regulatory processes, and/or the need to redistribute amino acid resources of the cell. Notably, transcripts encoding putative O₂-sensing proteins are

upregulated, indicating a possible mechanism by which algae sense and respond to the presence or absence of O_2 . Increased levels of several transcripts encoding proteins associated with anaerobic respiration are also observed. These proteins are potentially involved in pathways that compete with hydrogenase for reductant and could be the focus of future engineering efforts. A significant number of transcripts encoding proteins of unknown function are also observed to be differentially expressed.

Array data from distinct *C. reinhardtii* mutants, which are (1) unable to synthesize an active [FeFe]-hydrogenase or (2) defective in starch synthesis, have been obtained and will be analyzed. Wild-type and mutant strains will also be examined using cultures deprived of sulfate, a physiological condition that induces anaerobiosis in the light and results in sustained H_2 photoproduction. Finally, insights obtained from the cellular-metabolism and gene-expression data are being integrated into a larger systems framework that is focused on understanding the flexibility of whole-cell metabolism under rapidly changing environmental conditions. Moreover, this information may potentially be leveraged into metabolic-engineering strategies designed to optimize the production of fermentative products including H_2 and/or ethanol.

93 ^{MEWG}

Perspectives in Metabolic Flux Mapping

Jacqueline V. Shanks* (jshanks@iastate.edu)

Department of Chemical and Biological Engineering, Iowa State University, Ames, Iowa

Project Goals: Experimentally determine flux maps for plant systems (NSF) and microbes (Interagency Program on Metabolic Engineering).

Metabolic flux maps provide a quantitative depiction of carbon flow through competing metabolic pathways, thus providing: analysis of substrate utilization and product formation; flexibility or rigidity of carbon flow at network nodes; the rate of a given enzymatic reaction in vivo; and inferred availability of NADPH or ATP. Thus, metabolic fluxes are an important physiological characteristic complementary to levels of transcripts, proteins, and metabolites. The system-wide quantification of intracellular fluxes in an organism is called metabolic flux analysis (MFA). The most basic approach to MFA is stoichiometric MFA, which involves writing balances for intracellular metabolites based on the stoichiometry of the biochemical reactions in the metabolic network. This results in a system of linear equations, which are solved by employing extracellular and biomass synthesis flux measurements to resolve some or all degrees of freedom. Genome-wide or in silico flux models provide the solution space of feasible fluxes resulting from optimization of the balances to a global cellular goal, such as maximum growth rate. Recently, constraints to the in silico models provided by data from ^{13}C labeling experiments, have narrowed the solution space.

^{13}C metabolic flux analysis (^{13}C MFA), with isotope detection via GC/MS or NMR of metabolites (e.g. amino acids from hydrolyzed protein), quantifies intracellular metabolic fluxes for smaller reaction networks, where the fluxes are completely determined (in contrast to the in silico models). ^{13}C MFA provides redundant measurements for flux quantification, as well as testing the consistency of the network topology for the physiological conditions. Isotopomers, which are isomers of a metabolite that differ in the labeling state (^{13}C or ^{12}C) of their individual carbon atoms, are a central concept in the analysis and mathematical modeling of ^{13}C MFA.

Increasing levels of information can be obtained from ^{13}C labeling data when coupled with a stoichiometric model of the biochemical pathways and computational methods to solve for flux data in the smaller network. More rigorous analysis is indicated as one moves from analytical (a few flux ratios at metabolic branchpoints) to ^{13}C constrained flux analysis (stoichiometric model with a few flux ratios as constraints) to fully integrated determination of fluxes from all the experimental data and the stoichiometric and isotopomer balances. Iterative methods have been used to solve the full relationship of isotopomer balances and the NMR or GC measurements to provide consistency, and routines to minimize error from the overdetermined data sets are required.

^{13}C MFA studies of aerobic glycolysis in microorganisms have become “higher throughput” since simplifications to the metabolic network can be made, and ^{13}C constrained flux analysis can be used. For alternative physiological conditions, for example where anapleurotic pathways are active, reversibility of reactions are indicated, or substrates other than glucose are used, the development of a consistent network topology and the strategy for the choice of the label to obtain identifiable fluxes are not as straight-forward. Furthermore, due to compartmentation and the existence of parallel pathways in plants, more experimental measurements are needed than in microbial systems, and the number of isotopomer balances increases, further increasing the computational burden. As a note, to date in silico models for plants have not been developed. Thus, at this point, these more challenging systems are not yet ready for “high-throughput” measurements. However, a growing knowledge base in ^{13}C MFA in these systems should enable movement towards more genome-wide flux estimation. This presentation will summarize these points with approaches from our laboratory in determining ^{13}C -based metabolic flux maps in plants and microbes.

94 GTL

High-Resolution Functional Assignments of Genes through Mapping KEGG Pathways to Bacterial Genomes

Fenglou Mao, Hongwei Wu, and **Ying Xu*** (xyn@bmb.uga.edu)

Department of Biochemistry and Molecular Biology and Institute of Bioinformatics, University of Georgia, Athens, Georgia

Project Goals: Modeling of biological pathways and networks.

We have developed a computational capability for mapping KEGG metabolic pathways to sequenced bacterial genomes. This capability assigns genes of a bacterial genome to specific enzymatic roles of a given KEGG pathway using a two-level strategy: (a) initial assignment is based on the premise that a bacterial metabolic pathway is in general encoded by a number of (in general transcriptionally co-regulated) operons and based on predicted functions of individual genes possibly at a low-resolution level; and (b) filling the gaps, the unassigned enzymes, in a partially-assigned pathway based on the detected co-evolutionary, co-occurrence and co-regulated relationships and predicted protein-protein interactions between un-annotated genes and genes already assigned to the pathway. To facilitate automated functional assignment of genes, we have developed a number of supporting computational tools, including prediction of operons [1], uber-operons [2] and regulons (unpublished results).

A. Initial KEGG pathway mapping: We have developed a computational algorithm for mapping a KEGG pathway to a specified bacterial genome. The algorithm starts by searching each gene in the target genome against gene databases with annotated functions such as the *nr* database and making

functional predictions, possibly at a low-resolution level, based on identified homology relationship. Then a number of genes (possibly zero) with annotated functions will be predicted as possible candidates for each enzyme in the KEGG pathway, based on the match between the predicted gene functions and the enzyme. We then assign at most one candidate gene to each enzyme of the KEGG pathway, using the following criteria: the overall consistency between the predicted gene functions and their assigned enzymatic roles should be as high as possible, and the selected genes should be clustered as much as possible as the predicted operons. This problem is formulated as a constrained optimization problem, specifically a linear integer programming problem, and solved using a commercial linear integer programming solver COIN. This overall prediction capability has been implemented as a computer program, called PMAP-KEGG. Using this capability, we have mapped over 140 KEGG pathways to 300+ sequenced bacterial genomes, including *E. coli*, for which detailed validation has been done using pieces of information from multiple sources. For every sequenced bacterial genome, our mapping results cover a substantial fraction of all the genes in that genome. Detailed data will be reported in an extended version of this abstract.

It should be noted that the operon prediction for each target bacterial genome is made using three prediction programs, JPOP [1], OFS [3] and VIMSS [4]. A simple majority-vote scheme is used for the final operon prediction. In the actual formulation of the problem, we have also taken into consideration the predicted uber-operon information using our own prediction program [2], where a *uber-operon* represents a group of operons whose union is conserved across multiple genomes, which gives a higher prediction coverage than using operons alone.

B. Filling gaps in a partially assigned pathway: The mapped KEGG pathways often contain “gaps”, unassigned enzymatic roles, due to various reasons. We have developed a computational procedure attempting to fill in these gaps, using three types of information: (a) co-evolutionary and co-occurrence information between assigned genes and unassigned & un-annotated genes, (b) predicted regulon information (i.e., transcriptionally co-regulated operons), and (c) protein-protein interaction information derived using various techniques such as the gene fusion method. It has been generally known that co-evolutionary, co-occurrence and co-regulation information of genes can help to predict functional relatedness among genes, even when functions of some of the genes are unknown. By employing this idea, we have recently developed a computational technique for predicting genes that are possibly working closely together in the same biological process [5,6]. Using this capability, we have predicted an initial set of candidate genes for each “gap” in a partially assigned KEGG pathway. We have then predicted protein-protein interaction relationships between the candidate genes for each “gap” with the genes already assigned to the network neighborhoods of the gap. Our final prediction for each gap is selected, using a trained neural network, based on the predicted functional relatedness and protein-protein interaction. We found that we were able to make correct gene assignments (as top assignments), for about 30% of the gaps, on a large test set using well characterized *E. coli* pathways after manually removing some of the assigned genes (1-3 genes are randomly removed from each assigned pathway). Detailed results will be reported in an extended version of this abstract.

Concluding remarks: By assigning genes of a bacterial genome to KEGG pathways, we can provide functional prediction of genes at a high-resolution level (knowing exactly the functional role in a well understood metabolic pathway), compared to the low-resolution functional annotation typically provided by a genome annotation system, e.g., gene A encodes a protease, and also can assign un-annotated genes to possible functional roles in a metabolic pathway. Our computational prediction program consists of a number of prediction and analysis tools, which are pipelined together to facilitate large-scale applications. A database containing all mapped KEGG pathways to each of the 300+ sequenced bacterial genomes is currently being developed, and will be made publicly available within a few months. This collection of mapped pathways has provided a very rich set of information

for studies of bacterial metabolic pathways and their evolution. For example, by comparing the same mapped KEGG pathways across multiple genomes, we can derive information about how a pathway has evolved in adaptation to an organism's living environments, leading to general information about pathway evolution and adaptation.

Acknowledgement: This work was supported in part by the by National Science Foundation (NSF/DBI-0354771, NSF/ITR-IIS-0407204, NSF/CCF-0621700, NSF/DBI-0542119) and the US Department of Energy's Genomics: GTL Program.

References

1. X. Chen, Z. Su, P. Dam, B. Palenik, Ying Xu and T. Jiang, Operon prediction by comparative genomics: an application to the *Synechococcus WH8102* genome, *Nucleic Acids Research*, 32 (7), 2147 – 2157, 2004.
2. D. Che, G Li, F. Mao, H Wu, and Ying Xu, "Detecting uber-operons in microbial genomes", *Nucleic Acids Research*, 2006 (in press).
3. BP. Westover, JD. Buhler, JL. Sonnenburg, and J.I. Gordon, Operon prediction without a training set, *Bioinformatics*, 21, 880-888, 2005.
4. MN. Price, KH. Huang, EJ. Alm and AP. Arkin, A.P. A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res*, 33, 880-892, 2005.
5. F. Mao, Z. Su, V. Olman, P. Dam, Z. Liu and Ying Xu, Mapping of Orthologous Genes in the Context of Biological Pathways: an Application of Integer Programming, *Proc Natl Acad Sci USA*, 103, 129-134, 2006.
6. H. Wu, Z. Su, V Olamn, Ying Xu, Prediction of functional modules through comparative genome analysis and application of gene ontology, *Nucleic Acids Research*, 33, 2822-2837, 2005.

Section 3

Regulatory Processes

95 ^{GTL}

A Systems Approach to Characterizing Evolutionarily Conserved Transcriptional Complexes Elucidates the Architecture of a Global Regulatory Network in Archaea

Marc T. Facciotti*, David J. Reiss, Min Pan, Amardeep Kaur, Madhavi Vuthoori, Richard Bonneau, Paul Shannon, Alok Srivastava, Samuel M. Donahoe, Leroy Hood, and **Nitin S. Baliga** (nbaliga@systemsbiology.org)

Institute for Systems Biology, Seattle, Washington

Project Goals: MAGGIE: To characterize evolutionarily conserved protein complexes from a systems perspective

MAGGIE Component 3

Cells responding to dramatic environmental changes or undergoing a developmental switch typically change the expression of numerous genes. In bacteria sigma factors regulate much of this process

while in eukaryotes four RNA polymerases and a multiplicity of generalized transcription factors (GTFs) are required. Here, using a systems approach, we demonstrate how archaeal organisms accomplish similar large scale transcriptional segregation and modulation of related physiological functions with an expanded family of Transcription Factor B (TFB) proteins. Further, our data suggest that a gene regulatory circuit assembled through an evolution of protein-protein and protein-promoter interactions among the seven TFBs might mediate coordination of their regulatory functions. The findings reported here represent a significant contribution towards closing the gap in our understanding of gene regulation by GTFs for all three domains of life.

96 ^{GTL}

CRP and cAMP Regulatory Networks of *Shewanella oneidensis* MR-1 Involved in Anaerobic Energy Metabolism

Daad A. Saffarini,^{1*} Sheetal Shirodkar,¹ Yang Zhang,² and **Alexander S. Beliaev**²

¹University of Wisconsin, Milwaukee, Wisconsin and ²Pacific Northwest National Laboratory, Richland, Washington

Project Goals: This project is a component of the *Shewanella* Federation and as such contributes to the overall goal of applying the tools of genomics, leveraging the availability of genome sequence for 18 additional strains of *Shewanella*, to better understand the ecophysiology and speciation of respiratory-versatile members of this important genus. To understand these systems the SF is using genome-based approaches to investigate *Shewanella* as a system of integrated networks; first describing key cellular subsystems those involved in signal transduction, regulation, and metabolism—then building towards understanding the function of whole cells and, eventually, cells within populations. As a general approach, the SF is collectively employing complimentary top-down bioinformatics-based genome functional predictions, high-throughput expression analyses, and functional genomics approaches to uncover key genes as well as metabolic and regulatory networks. The bottom-up component employs more traditional approaches including genetics, physiology and biochemistry to test or verify predictions. This information will ultimately be linked to analyses of signal transduction and transcriptional regulatory systems and used to develop a linked model that will contribute to understanding the ecophysiology of *Shewanella* in redox stratified environments.

Shewanella oneidensis MR-1 is a facultative metal-reducing bacterium with extensive respiratory versatility. Unlike many bacteria studied to date, the ability of *S. oneidensis* to grow anaerobically with several electron acceptors is regulated by the cAMP-receptor protein (CRP) and not the fumarate-nitrate reduction (FNR) regulator. CRP-deficient mutants of MR-1 are impaired in anaerobic respiration and growth with Fe(III), Mn(IV), fumarate, nitrate, and DMSO due to the loss of anaerobic terminal reductases as well as heme and porphyrin biosynthesis deficiencies.

Genetic and biochemical evidence confirms that similarly to other bacteria, CRP in *S. oneidensis* is activated upon binding to cAMP. The genome sequence of *S. oneidensis* contains three genes that are predicted to encode adenylate cyclases. This prediction was confirmed by complementation of *E. coli* mutants that lack the adenylate cyclase gene *cyaA*. An *S. oneidensis* triple mutant that lacks the adenylate cyclase genes (*cyaA*, *cyaB*, and *cyaC*) was generated and found to be deficient in anaerobic respiration similar to the *crp* mutant. To further elucidate the role of CRP and to understand the

mechanisms of cAMP-dependent gene expression under anaerobic conditions in *S. oneidensis*, a combination of experimental and computational approaches have been applied.

Expression profiling of mRNA levels suggests that CRP regulates gene expression directly and indirectly. Global transcriptome comparisons of the wild-type *vs.* the Δcrp mutant indicate that CRP positively regulates the expression of genes involved in anaerobic energy generation and transcriptional regulation. These include the anaerobic DMSO reductase (*dmsAB*), the decaheme *c*-type cytochrome cluster (*omcA*, *mtrCAB*), the anaerobic formate reductase (*fdhABC*), the pyruvate-formate lyase (*pflAB*), and genes encoding the two-component signal transduction involved in anaerobic respiration of sulfur compounds. Mobility shift assays using purified CRP and DNA upstream of the DMSO reductase operon, indicate that this protein directly regulates the expression of the *dms* genes. Regulation of other anaerobic reductase genes appears to involve additional proteins that are under CRP control. Recent experiments identified a two-component signal transduction system (SO4155 and SO4157) that is upregulated by CRP and that regulates the expression of thiosulfate reductase genes. Additionally, we identified a regulatory protein (SO0490) that appears to negatively regulate sulfur reduction. SO0490 is regulated by CRP as suggested by microarray analysis and transcriptional *lacZ* fusions. Our results suggest that a complex regulatory network, with CRP as the global regulator, controls anaerobic respiration in *S. oneidensis*. Further work is underway to further elucidate the mechanisms of anaerobic gene regulation in *S. oneidensis* MR-1.

97 ^{GTL}

Mapping the Genome-Scale Regulatory Network of *Shewanella oneidensis* MR-1: Identification of Metal-Respiratory Regulation

M.E. Driscoll, F.S. Juhn, J.J. Faith, B. Hayete, J.J. Collins, and **T.S. Gardner*** (tgardner@bu.edu)

Department of Biomedical Engineering, Boston University, Boston, Massachusetts

Project Goals: To map the transcriptional regulatory pathways underlying metal respiration and carbon source utilization in *Shewanella oneidensis* MR-1

Shewanella's respiratory versatility reflects its diverse environmental ecology. In recent years, *Shewanella* species have been isolated from fresh water lake sediments, surface waters of the Sargasso sea, hydrothermal vents of the deep Pacific, marine sediments from around the globe, mollusks and spoiling fish. To successfully compete across these distinct niches, many of which represent dynamically shifting redox environments, *Shewanella* species must be respiratory generalists. In particular, *Shewanella*'s capacity for driving respiration with metals as electron acceptors – including arsenic and uranium – has made it a candidate for use in microbial fuel cells and environmental remediation applications.

Though multiple *Shewanella* genomes have been sequenced and many of the enzymes involved in electron transport have been identified, little is known about how this metabolic machinery is regulated. To this end, we have designed the first high-density oligonucleotide array for *S. oneidensis* MR-1 to observe and model its global gene expression. We have profiled gene expression in more than one hundred environmental conditions which vary carbon sources, electron acceptors, and environmental factors within physiological ranges. These conditions represent the first phase of more than 300 planned conditions.

Using this initial expression data, we have predicted a regulatory network of more than 200 transcriptional interactions for *S. oneidensis* using the CLR algorithm we recently developed. The CLR algorithm, a novel extension of the relevance networks class of algorithms, has been successfully validated in *Escherichia coli* for mapping global regulatory networks.¹ In the *E. coli* study, 741 novel regulatory interactions were identified at a 60% true positive rate.

The predicted *S. oneidensis* regulatory map suggests several novel relationships between as-yet uncharacterized transcription factors and genes governing heme synthesis and cytochromes implicated in iron and manganese reduction. Analysis of our expression profiles also suggests that *S. oneidensis* possesses a broader capacity for carbon source utilization than has been previously observed.

While electron acceptor pathways have been a dominant focus of study for *S. oneidensis* to date, bacterial respiration involves a complex interplay between both electron donor and acceptor pathways. A deeper knowledge of both electron donor and acceptor metabolism is relevant not only to understanding the role of *S. oneidensis* in its natural environments, but also towards the optimization of dissimilatory metal reducing bacteria for multiple applications.

Reference

1. JJ Faith, B Hayete, JT Thaden, I Mogno, J Wierzbowski, G Cottarel, S Kasif, JJ Collins, TS Gardner. (2007) Large-scale mapping and validation of *Escherichia coli* transcriptional regulatory interactions from a compendium of expression profiles. *PLoS Biology*. 5(1): e8.

98 ^{GT}

A Web-Based Tool for Visualizing *Shewanella* Gene Expression Profiles in Their Chromosomal Context

J.J. Faith,^{1*} R. Sachidanandam,² and T.S. Gardner¹ (tgardner@bu.edu)

¹Department of Biomedical Engineering, Boston University, Boston, Massachusetts and ²Cold Spring Harbor Laboratory, Cold Spring Harbor, New York

Project Goals: To map the transcriptional regulatory pathways underlying metal respiration and carbon source utilization in *Shewanella oneidensis* MR-1

Common microarray analysis procedures yield lists of genes, whose expression changes significantly in response to an environmental or genetic perturbation. The functional role for most of these expression changes is typically unknown, and the often large number of changed genes hinders human interpretation of their role. In bacteria, genes with similar functional roles often exhibit chromosomal proximity and therefore operate as a coexpressed module, even when part of distinct operons. Moreover, the expression of an RNA in intergenic regions can also suggest a probable role as noncoding regulatory RNA. To facilitate the sharing, discovery and analysis of *Shewanella oneidensis* expression data and gene function, we developed a web-based genome browser where users can dynamically choose any two sets of experiments from the current *Shewanella oneidensis* Affymetrix microarray compendium and view expression levels or changes of genes in their chromosomal context. This capability is built into the M3D Database of Affymetrix microarray compendia. M3D includes compendia of several hundred expression profiles for multiple microbes including *S. oneidensis*, *E. coli*, and *S. cerevisiae*, and provides viewing and raw data download tools.

Reference

1. Faith JJ, Fusaro V, Driscoll M, Juhn F, Cosgrove E, Hayete B, Gardner TS. Many Microbes Microarray Database (M3D). <http://m3d.bu.edu>

99 ^{GTL}Comparative Genomics of Signal Transduction in *Shewanella*

Luke E. Ulrich and **Igor B. Zhulin*** (joulineib@ornl.gov)

Joint Institute for Computational Sciences, University of Tennessee, Knoxville, Tennessee and Oak Ridge National Laboratory, Oak Ridge, Tennessee

Project Goals: This project is a component of the *Shewanella* Federation and as such contributes to the overall goal of applying the tools of genomics, leveraging the availability of genome sequence for 18 additional strains of *Shewanella*, to better understand the ecophysiology and speciation of respiratory-versatile members of this important genus. To understand these systems the SF is using genome-based approaches to investigate *Shewanella* as a system of integrated networks; first describing key cellular subsystems—those involved in signal transduction, regulation, and metabolism—then building towards understanding the function of whole cells and, eventually, cells within populations. As a general approach, the SF is collectively employing complimentary “top-down” -bioinformatics-based genome functional predictions, high-throughput expression analyses, and functional genomics approaches to uncover key genes as well as metabolic and regulatory networks. The “bottom-up” component employs more traditional approaches including genetics, physiology and biochemistry to test or verify predictions. This information will ultimately be linked to analyses of signal transduction and transcriptional regulatory systems and used to develop a linked model that will contribute to understanding the ecophysiology of *Shewanella* in redox stratified environments

The availability of genome sequence of 13 *Shewanella* strains provided us with a unique opportunity to unravel the evolutionary trends of signal transduction systems on a single genus scale. We used recently developed MiST database to analyze signal transduction profiles in *Shewanella* [1, 2]. Results obtained allowed us to link the overall signal transduction profile of a given strain to its metabolic potential. We also found that the distinct natural history of signal transduction proteins in *Shewanella* species provides useful markers for improving their taxonomy. Overall, all *Shewanella* strains maintain a similar profile of signal transduction with respect to protein families that constitute regulatory networks; however, the number of proteins in each family varies significant from strain to strain. The most dramatic changes in the overall composition of signal transduction is observed in *S. denitrificans* OS217. This strain has only 148 transcriptional regulators, whereas the strain *S. putrefaciens* CN-32, which has a comparable genome size, has 181 transcriptional regulators. Adjusted to the genome size, *S. denitrificans* OS217 is the outlier in each major category of signal transduction. Such reduction in this important functional category correlates with the lack of most of the anaerobic respiratory machinery that is present in other strains. This finding strongly supports the hypothesis that many regulatory pathways in *Shewanella* control its branched electron transport system.

Recent divergence of the 13 *Shewanella* strains allowed us to identify distinct cases of horizontal gene transfer and strain-specific gene loss that are most common trends in the evolution of regulatory systems. These findings lead to a better resolution of *Shewanella* taxonomy

References

1. Ulrich, L.E., E.V. Koonin, and I.B. Zhulin (2005) One-component systems dominate signal transduction in prokaryotes. *Trends Microbiol* 13: 52-56.
2. Ulrich, L.E., and I.B. Zhulin (2006) MiST: a microbial signal transduction database. *Nucleic Acids Res.* Nov 28, [Epub ahead of print].

100 ^{GT}

Comparative Genomics of Transcriptional Regulation of Metabolic Pathways in *Shewanella* Species

Dmitry Rodionov^{1*} (rodionov@burnham.org), Mikhail Gelfand,² Margaret Romine,³ and **Andrei Osterman**^{1,4}

¹Burnham Institute for Medical Research, La Jolla, California; ²Institute for Information Transmission Problems RAS, Moscow, Russia; ³Pacific Northwest National Laboratory, Richland, Washington; and ⁴Fellowship for Interpretation of Genomes, Burr Ridge, Illinois

Project Goals: This project is a component of the *Shewanella* Federation and as such contributes to the overall goal of applying the tools of genomics, leveraging the availability of genome sequence for 18 additional strains of *Shewanella*, to better understand the ecophysiology and speciation of respiratory-versatile members of this important genus. To understand these systems the SF is using genome-based approaches to investigate *Shewanella* as a system of integrated networks; first describing key cellular subsystems—those involved in signal transduction, regulation, and metabolism—then building towards understanding the function of whole cells and, eventually, cells within populations. As a general approach, the SF is collectively employing complementary “top-down” - bioinformatics-based genome functional predictions, high-throughput expression analyses, and functional genomics approaches to uncover key genes as well as metabolic and regulatory networks. The “bottom-up” component employs more traditional approaches including genetics, physiology and biochemistry to test or verify predictions. This information will ultimately be linked to analyses of signal transduction and transcriptional regulatory systems and used to develop a linked model that will contribute to understanding the ecophysiology of *Shewanella* in redox stratified environments.

Genomics-based reconstruction and comparative analysis of regulons have been utilized to predict transcriptional regulatory subnetworks in a number of model organisms (e.g. *E. coli*) providing an extremely useful resource for interpreting results from microarray datasets and deriving genome-wide regulatory models. Our approach combines the detection of transcription factors binding sites and cross-genome comparison with the analysis of the genomic and functional context inferred by metabolic reconstruction. The recent availability of genome sequences for 11 *Shewanella* species allowed us to perform, for the first time, a detailed comparative analysis of transcriptional regulons for the key pathways involved in central metabolism, production of energy and biomass, metal ion homeostasis and stress response. This analysis provides insights not only into regulatory networks in *S. oneidensis* MR-1 (our model organism) but also in the *Shewanella* lineage as a whole. In addition, this approach allows us to improve gene function and metabolic pathway assignments, as well as to accurately predict functions of previously uncharacterized gene families. Results of these studies will serve as an important shared resource of the *Shewanella* Federation research team who together seek to develop a systems-level understanding of *Shewanellae* metabolic and regulatory networks.

Using this approach, we identified candidate binding sites for more than 20 transcriptional factors of known specificity, including global regulators (CRP, FNR, ArcA, Fur) and specialized regulators of the metabolism of nitrogen (NarP, IscR, NsrR, DNR, NorR), vitamins and amino acids (BirA, ArgR, MetJ, TyrR, HutC), fatty acids (FadR, FabR), and sugars (NagR, SdaR, ScrR, GntR), as well as of the central carbon metabolism (PdhR, HexR), Fe-S cluster assembly (IscR) and ribonucleotide reduction (NrdR). In addition, we identified candidate binding sites for previously uncharacterized regulators, termed NagR (SO3516), IlvR (SO1898) and FadQ (SO2493) tentatively implicated in the control of the metabolism of N-acetylglucosamine, branch chain amino acids, and fatty acids, respectively.

Cross-validation of these predictions with the results of microarray analysis combined with targeted gene knock-outs are currently under way. The first results reveal substantial consistency between our predictions and gene expression profiles, as will be illustrated for the arginine and N-acetylglucosamine regulons. We have also mapped the genes and operons controlled by five types of metabolite-binding riboswitches (*B12*, *LYS*, *RFN*, *THI*, *GLY*), and six translational attenuators of amino acid biosynthesis pathways (*ilv*, *leu*, *his*, *thr*, *trp*, *phe* operons). Although some diversity of the predicted regulons is observed within the collection of *Shewanella* spp., the most striking difference in the overall regulatory strategy is revealed by comparison with *E. coli* and other gamma-proteobacteria. Multiple examples of divergence and adaptive evolution of regulatory networks were detected and include regulon “shrinking” (as in case of FadR), “expansion” (as in case of PdhR and HexR), “mergers”, and “split-ups”, as well as multiple cases of using nonorthologous regulators to control equivalent pathways or orthologous regulators to control distinct pathways. Among the most notable are the differences in the regulon content and a respective role of global regulators, such as CRP. These and other observations, indicate that many aspects of metabolic regulation in *Shewanella* species., are substantially different from regulatory network models that were largely derived from studies in *E. coli*.

101 MEWG

Biological Aspects of Deciphering and Engineering Regulatory Networks

George N. Bennett* (gbennett@bioc.rice.edu)

Department of Biochemistry and Cell Biology, Rice University, Houston, Texas

Project Goals: Studies to enhance chemical production in microbes.

In efforts to detect and modify regulation of pathways for a particular goal there are strategies that can be implemented in two circumstances, one the synthetic approach where major components of the pathway are known and the appropriate regulation of the various enzymes can be adjusted by taking advantage of modeling and combinatorial assembly methods; and the other if the proteins to be altered are not obvious and thus the modification must take a more empirical approach with selection or screening methods being the issues.

In known expression systems, it is still needed to detect levels of regulation, for example determination of the level of functioning of a protein (enzyme for metabolic process) more than knowledge of the level of gene expression is needed to understand how the activity varies under the physiological condition contemplated for use. This is the situation if genes from various sources are placed together in a new way to form a non-endogenous pathway as often is the case in metabolic engineering. Detection of proteins, metabolites etc from systems biology approach of measurements in different

genetically engineered cells under various conditions can help in this endeavor. Due to the large number of possible combinations of mutations and conditions a way is needed to minimize or focus the experimental measurements on the most appropriate ones to examine. Such experiments should make an effort to take into account the effects of intercellular conditions produced by introduced changes on related protein activities (other enzymes of pathway, regulatory factors, functional state of activity of the enzyme or regulatory proteins) and models that include this interaction information would be more comprehensive.

In order to carry out appropriate modification of regulation there are relatively straightforward approaches in the case of desired changes in specific known expression controls such as through modification of transcriptional events and to a lesser extent, modification of more general aspects of cellular physiology (redox and air, enzyme stability, enzyme parameters, osmotic conditions). To target regulation to specific pathway, designed regulation of small units can be employed. These involve the use of known regulated promoters that are varied to adjust the level of constitutive expression, and can be combined with terminators or RNA structural elements to afford variation of level of expression. In the case of less known processes found in many industrial organisms that have less genetic and biochemical literature, efforts are more a matter of perturbing a somewhat more global functioning system and screening or selection for those altered cells that perform better, then analyzing and combining the most promising. The idea of eliminating complicating or undesired processes that may obscure the regulation you would like to enhance is an useful experimental strategy. For these wider scope effects, or regulation of unknown factors with less obvious connections to the metabolically engineered process, modification of transcription factors such as sigma factors, Zn-finger motif factors, general or global transcription factors may be used in combination with powerful selection or screening systems for the desired property.

102 ^{GTL}

Characterization of Behavioral Responses in *Shewanella oneidensis*

Jun Li,¹ Margie Romine,² and **Mandy Ward**^{1*} (mjward@jhu.edu)

¹Department of Geography and Environmental Engineering, Johns Hopkins University, Baltimore, Maryland and ²Pacific Northwest National Laboratory, Richland, Washington

Project Goals: This project is a component of the *Shewanella* Federation and as such contributes to the overall goal of applying the tools of genomics, leveraging the availability of genome sequence for 18 additional strains of *Shewanella*, to better understand the ecophysiology and speciation of respiratory-versatile members of this important genus. To understand these systems the SF is using genome-based approaches to investigate *Shewanella* as a system of integrated networks; first describing key cellular subsystems—those involved in signal transduction, regulation, and metabolism—then building towards understanding the function of whole cells and, eventually, cells within populations. As a general approach, the SF is collectively employing complimentary “top-down” - bioinformatics-based genome functional predictions, high-throughput expression analyses, and functional genomics approaches to uncover key genes as well as metabolic and regulatory networks. The “bottom-up” component employs more traditional approaches including genetics, physiology and biochemistry to test or verify predictions. This information will ultimately be linked to analyses of signal transduction and transcriptional regulatory systems and used to develop a linked model that will contribute to understanding the ecophysiology of *Shewanella* in redox stratified environments.

Shewanella species are often found in redox stratified environments. This, and the ability of the genus to utilize many different terminal electron acceptors for anaerobic respiration, suggests that sensing, and responding behaviorally to, different redox conditions may be integral to the lifestyle of these microorganisms. The first studies on chemotaxis by *Shewanella oneidensis* MR-1 were conducted over ten years ago. These studies revealed that, unlike *Escherichia coli*, MR-1 does not respond to a number of carbon sources including methanol, ethanol, acetate, lactate, propionate, pyruvate, citrate, several amino acids (both individually and in combinations with vitamins), malate, glucose, yeast extract, and peptone. Formate did produce a weak response, although only under anaerobic conditions. Instead, chemotaxis to several respiratory electron acceptors (nitrate, nitrite, fumarate, TMAO, DMSO, and to a lesser extent thiosulfate) was observed. The background presence of certain energetically favorable and energetically less favorable electron acceptors (O_2 , nitrate, nitrite, DMSO, and elemental sulfur) inhibited these responses, although they did not necessarily inhibit reduction of the test electron acceptor. The genome sequence of MR-1 was released seven years later, providing us with a valuable new resource for revisiting this earlier data and designing new experiments targeted at developing a more complete understanding of tactic responses in MR-1. The sequence of MR-1, as well as those soon to become available for 18 other *Shewanella* genomes, allows us to compare domain architecture of the signal transduction proteins in this genus (presented by Dr. Jhoulin) and to conduct genome neighborhood analysis (presented by Dr. Osterman) to identify candidate metabolisms that may illicit chemotactic responses. Furthermore, high throughput analyses (global proteomics and microarrays) available from other collaborators within the *Shewanella* Federation will provide additional clues for designing experiments to investigate behavioral responses. In this presentation, we will provide preliminary findings derived from characterization of various mutants defective in tactic or respiratory functions as well as preliminary insights gleaned from surveying available microarray and proteome data from MR-1.

The genome sequence revealed three clusters of chemotaxis genes, each of which could potentially encode a complete chemotactic signal transduction pathway. Each locus encodes a CheY, CheA, CheW, CheR, and CheB protein. However, two genes (*cheA-2* and an *mcp*) in the predicted Cluster II operon are interrupted by a transposase, indicating that this entire signal transduction pathway may be degenerate. The cluster III locus encodes the sole CheZ protein and the entire subsystem required for biosynthesis and assembly of the flagellum, as well as the sigma factors σ^{54} and σ^{27} , and an anti- σ^{28} factor. Examination of our proteome data revealed that the proteins associated with cluster III are by far the most frequently observed across all our datasets. Consequently, the cluster III-encoded proteins probably constitute the main chemotactic signal transduction system in MR-1.

In-frame deletions of the *cheA-1* and *cheA-3* genes have been constructed and the phenotypes analyzed in swarm plate assays. These experiments showed that the Δ *cheA-3* mutant is non-chemotactic to a range of anaerobic electron acceptors, supporting the previous prediction that the cluster III genes encode the main chemotactic signal transduction pathway in *S. oneidensis* MR-1. Motion analysis of this mutant showed that it has the smooth swimming phenotype associated with non-chemotactic mutants. The Δ *cheA-1* mutant showed a reduced-swarming phenotype, suggesting that the cluster I-encoded signal transduction pathway is not essential for chemotaxis to anaerobic electron acceptors, but may play a role in optimizing responses.

A total of 29 different methyl-accepting chemotaxis proteins (MCPs) are encoded by the genome (2 by pseudogenes) suggesting that the variety of sensory inputs detected by *S. oneidensis* MR-1 is far greater than has presently been discovered. Data from DNA microarray and proteomic analyses of MR-1 have provided insights into the range of environmental information that these MCPs may sense. For example, induction of specific *mcp* genes in response to chromium and uranium (1), acidic pH (2), and iron and manganese (3) suggest that behavioral responses to these stimuli warrant inves-

tigation. Mutants with insertions in several *mcp* genes have been constructed and tested for loss of behavioral responses to anaerobic electron acceptors and metals. To date, the only mutants that show swarming defects to electron acceptors have insertions in the *mcp* genes that encode receptors with redox sensing PAS domains (SO0584, SO1385, SO2123, and SO3404). This result suggests that several MCPs are involved in energy taxis rather than chemotaxis. Two *c*-type cytochrome mutants display the same swarming defects as the aforementioned CheA-3 and PAS-MCP mutants, suggesting that energy taxis in *S. oneidensis* may involve respiratory electron transport sensing. Additional mutants are being constructed based on predictions from the DNA microarray and proteomics data, and new assays will be designed to test for loss of responses to specific stimuli.

The integration of predictions, based on large scale genomic / proteomic analyses, with single gene mutational approaches, has so far proven successful in determining roles for some of the Che and MCP proteins in *S. oneidensis* MR-1. Consequently, this approach will be continued and combined with comparative analyses that will make use of information available for other *Shewanella* species. The ability to design experiments based on genomics will facilitate the rapid generation of results that will be shared with other researchers in the *Shewanella* Federation by incorporation into the Integrated Knowledge Source (ORNL).

References

1. Bencheikh-Latmani *et al.* (2005) Global transcriptional profiling of *Shewanella oneidensis* MR-1 during Cr(VI) and U(VI) reduction. *Appl. Environ. Microbiol.*, **71**:7453-7460.
2. Leaphart *et al.* (2006). Transcriptome profiling of *Shewanella oneidensis* gene expression following exposure to acidic and alkaline pH. *J. Bacteriol.*, **188**:1633-1642.
3. Beliaev *et al.* (2006) Global transcriptome analysis of *Shewanella oneidensis* MR-1 exposed to different terminal electron acceptors. *J. Bacteriol.*, **187**:7138-7145.

103 ^{GTL}

Development of *in vitro* Transcription System using Recombinant *Shewanella oneidensis* RNA Polymerase

Younggyu Kim* (ykim@chem.ucla.edu), Sam On Ho, Natalie Gassman, and **Shimon Weiss** (sweiss@chem.ucla.edu)

Department of Chemistry and Biochemistry, University of California, Los Angeles, California

Project Goals: Overall objective: Characterization of gene regulation in *Shewanella oneidensis* MR-1 using single molecule fluorescence spectroscopy.

1. **In vitro reconstitution of MR-1 transcription machinery**
2. **Structural and functional characterization of the transcription system using single molecule fluorescence spectroscopy.**

Shewanella oneidensis MR-1 grows aerobically and anaerobically using a variety of electron acceptors including fumarate, nitrate, nitrite, thiosulfate, elemental sulfur, trimethylamine N-oxide (TMAO), dimethyl sulfoxide (DMSO) and metal pollutants[1]. It has gained significant research attention as a potential bioremediation tool due to its ability to process biohazardous pollutants. The complete genome was sequenced by the Institute for Genomic Research (TIGR; <http://www.tigr.org/>) under the support of the U.S. Department of Energy (Microbial Genome Program and NABIR Program)[2], enabling us to study the biology of the bacterium at a system-wide level.

One of our main goals is the elucidation of gene regulation in MR-1. It is a complex process accompanied with a network of biomolecular interactions that affect the level of gene expression. Transcription is the first step in gene expression and is the step at which most of the gene regulation takes place. RNA polymerase (RNAP) is responsible for the transcription and is, directly or indirectly, a target for the most of gene regulation. Therefore, understanding the basic transcription mechanism is an important step for the study. The core enzyme of bacterial RNA polymerase consists of five subunits: α dimer (α_2), β , β' , and ω . When the core enzyme binds to one of its initiation factors (σ s), a holoenzyme is formed and it is then capable of initiating a promoter-specific transcription. Although the mechanism has been very well studied in *E. coli*, it is poorly understood in the *Shewanella* species.

In order to characterize the basic transcription machinery of the MR-1, we designed an *E. coli* expression system, co-overexpression and *in vivo* assembly of MR-1 RNA polymerase subunits in *E. coli*. We created an *E. coli* expression construct that produces polycistronic mRNA containing all MR-1 RNAP subunit coding sequences (α , β , β' , and ω) from a single T7 promoter; the C-terminus of α subunit is fused to a 6 histidine tag. The RNAP was purified by IMAC (Immobilized Metal Affinity Chromatography) and further purified by anion exchange chromatography (MonoQ). The purified multisubunit RNAP core (~400 kDa) shows a correct stoichiometric subunit ratio. Although there is a significant similarity (80-90%) in the subunit coding sequences between MR-1 and *E. coli*, we found that the subunits are not interchangeable between the two species, indicating the presence of lineage-specific coding sequences.

The recombinant RNAP is functional: (i) the core RNAP drives promoter-independent transcription, a characteristic of most core RNAPs. (ii) it forms a holoenzyme with the initiation factor, σ^{70} , responsible for house-keeping gene transcription; (iii) electrophoretic mobility shift assay shows that the holo RNAP binds to a target promoter specifically and forms an open complex; and (iv) the open complex initiates promoter-specific transcription. It is also capable of interacting with transcriptional activators. For example, MR-1 CAP (catabolite activator protein), one of transcription activators, interacts with the RNAP on the DMSO reductase promoter[3] containing potential CAP binding site and increases transcription activity. The initial characterization was conducted on the basis of well-established *E. coli* system because the amino acid sequences of the RNAP are highly conserved throughout evolution. Nevertheless, we found another remarkable difference between two RNAPs. While *E. coli* RNAP has the highest transcriptional activity at 37°C, MR-1 RNAP exhibits the optimal activity below 30°C. It suggests that the difference in lineage-specific sequences accounts for the trait which matches with the MR-1's natural habitat.

This system can be utilized for biochemical and biophysical studies in gene regulation. Those include studies in: core RNAP with different initiation factors, RNAP-transcriptional regulator interaction, structural studies of the transcription machinery using a single molecule fluorescence spectroscopy[4], and mutational studies of the enzymes having alterations in lineage-specific sequences or in transcription activation (or repression)-responsive sequences. Therefore, the newly established *in vitro* transcription system will serve as an important tool in order to study the gene regulation in MR-1.

References

1. Venkateswaran, K., et al., *Polyphasic taxonomy of the genus Shewanella and description of Shewanella oneidensis sp. nov.* International journal of systematic bacteriology, 1999. **49 Pt 2**: p. 705-24.
2. Heidelberg, J.F., et al., *Genome sequence of the dissimilatory metal ion-reducing bacterium Shewanella oneidensis.* Nature biotechnology, 2002. **20**(11): p. 1118-23.
3. Saffarini, D.A., R. Schultz, and A. Beliaev, *Involvement of cyclic AMP (cAMP) and cAMP receptor protein in anaerobic respiration of Shewanella oneidensis.* Journal of bacteriology, 2003. **185**(12): p. 3668-71.
4. Kapanidis, A.N., E. Margeat, S.O. Ho, E. Kortkhonja, S. Weiss, and R.H. Ebright, *Initial transcription by RNA polymerase proceeds through a DNA-scrunching mechanism.* Science, 2006. **314**(5802): p. 1144-7.

104 ^{GTL}Genetic Analysis of Anaerobic Respiration in *Shewanella oneidensis* MR-1

Jizhong Zhou,¹ Haichun Gao,^{1,3*} Xiaohu Wang,⁵ Soumitra Barua,¹ Yunfeng Yang,² Samantha B. Reed,⁴ Dave Culley,⁴ Zamin Yang,² Christopher Hemme,¹ Zhili He,¹ Margaret Romine,⁴ Kenneth Nealson,⁶ James M. Tiejde,³ Timothy Palzkill,⁵ and James K. Fredrickson⁴

¹University of Oklahoma, Norman, Oklahoma; ²Oak Ridge National Laboratory, Oak Ridge, Tennessee; ³Michigan State University, East Lansing, Michigan; ⁴Pacific Northwest National Laboratory, Richland, Washington; ⁵Baylor College of Medicine, Houston, Texas; and ⁶University of Southern California, Los Angeles, California

Project Goals: This project is a component of the *Shewanella* Federation and as such contributes to the overall goal of applying the tools of genomics, leveraging the availability of genome sequence for 18 additional strains of *Shewanella*, to better understand the ecophysiology and speciation of respiratory-versatile members of this important genus. To understand these systems the SF is using genome-based approaches to investigate *Shewanella* as a system of integrated networks; first describing key cellular subsystems—those involved in signal transduction, regulation, and metabolism—then building towards understanding the function of whole cells and, eventually, cells within populations. As a general approach, the SF is collectively employing complimentary “top-down” -bioinformatics-based genome functional predictions, high-throughput expression analyses, and functional genomics approaches to uncover key genes as well as metabolic and regulatory networks. The “bottom-up” component employs more traditional approaches including genetics, physiology and biochemistry to test or verify predictions. This information will ultimately be linked to analyses of signal transduction and transcriptional regulatory systems and used to develop a linked model that will contribute to understanding the ecophysiology of *Shewanella* in redox stratified environments.

Shewanella oneidensis MR-1, a facultative γ -proteobacterium, possesses remarkably diverse respiratory capacities. In addition to aerobic respiration, *S. oneidensis* can anaerobically respire various organic and inorganic substrates, including fumarate, nitrate, nitrite, thiosulfate, elemental sulfur, trimethylamine N-oxide (TMAO), dimethyl sulfoxide (DMSO), Fe(III), Mn(III) and (IV), Cr(VI), and U(VI). However, the molecular mechanisms underlying the anaerobic respiratory versatility of MR-1 remain poorly understood. As a part of the *Shewanella* Federation efforts, we have used integrated genomic, proteomic and computational technologies to study the regulation of energy metabolism of this bacterium from a systems-level perspective.

ArcA. The Arc two-component system is a major control system for the regulation of many genes involved in aerobic/anaerobic respiration and fermentative metabolism in *Escherichia coli*. MR-1 genome contains a gene encoding a putative ArcA homolog with ~81% amino acid sequence identity to the *E. coli* ArcA protein but no full-length arcB gene homolog. Results from physiological, microarray and computational analyses of an arcA deletion mutant revealed that the regulon of *S. oneidensis* ArcA differs significantly from that of *E. coli*. For instance, ArcA does not appear to be involved in regulation of the TCA cycle in *S. oneidensis*. Among the 50 operons controlled by *S. oneidensis* ArcA regulon, only 6 operons were shared by *E. coli* ArcA.

To further probe the role of ArcA in *Shewanella*, electrophoretic motility shift assay (EMSA), DNase I footprinting, and LacZ reporter gene assays were conducted. As predicted, *Shewanella* and *E. coli* ArcA bind a similar 15bp DNA motif and phosphorylation of Asp54 residue is still essential for

Shewanella ArcA activation. Surprisingly, however, our experimental results suggest *Shewanella* ArcA is constitutively activated by phosphorylation under both aerobic and anaerobic growth conditions and that activated ArcA regulates genes involved in the anaplerotic *sfcA* shunt, H₂ metabolism and terminal DMSO (dimethylsulfoxide) reduction. These findings indicate that ArcA is not a major redox response regulator in *Shewanella*. The protein-promoter binding interactions were also examined using the promoter binding microarray (PBM) for the ArcA protein. Using this array, the contribution of each nucleotide to the binding of ArcA within the ArcA binding site was evaluated by point mutations. Results indicate that several nucleotides are essential for ArcA binding and that the flanking sequence in addition to the 15bp motif appears to play a role in ArcA binding to the promoter.

Nitrate respiration. *S. oneidensis* MR-1 is able to reduce nitrate to ammonium under anaerobic conditions. *napDAGHB* gene cluster encoding periplasmic nitrate reductase (*NapA* of the *Nap* system) and accessory proteins and an *nrfA* gene encoding periplasmic nitrite reductase (*NrfA* of the *Nrf* system) were identified from the *S. oneidensis* genome. The *Nap* system catalyzes respiratory reduction of nitrate to nitrite and the *Nrf* system converts nitrite to ammonium. However, the genome lacks both *napC* and *nrfH*, which are essential for reduction of nitrate to nitrite and nitrite to ammonium, respectively. Mutation in *napA* renders the cells incapable of reducing nitrate to nitrite under either aerobic or anaerobic conditions. Similarly, mutation in *nrfA* eliminates reduction of nitrite to ammonium. Furthermore, a strain carrying the deleted *napB* gene exhibits significant differences in nitrate reduction compared to both MR-1 and the *napA* mutant. A further analysis reveals that the mutation causes reduction of nitrate to ammonium without nitrite accumulation. Mutational analysis of a *napB-nrfA* double mutant indicates that the cells missing *napB* also employ *Nap* and *Nrf* systems for nitrate and nitrite reduction. In an attempt to identify candidate replacements for *NapC* and *NrfH* functions, both microarray and mutational analyses have been performed. The results suggest that *CymA* is likely to be the functional replacement of both *NapC* and *NrfH* and a conceptual model is proposed. It appears that *NapB* is the preferred electron transferring protein and has an absolute priority in accepting electron from *CymA*. In the presence of *NapB*, *NrfA* could not function efficiently due to the lack of electrons from *CymA*.

Functional analysis of *c*-type cytochromes. To investigate the role of *c*-type cytochrome genes in anaerobic respiration, targeted deletions of 37 out of 44 predicted intact *c*-type cytochrome encoding genes have been generated by either homologous cross-over using host-encoded recombinases or by introduced phage *cre-loxP* recombinases. Each mutant was tagged with unique bar codes (i.e. short synthetic oligonucleotides) to facilitate tracking of the individual strains in planned competitive growth studies to determine the fitness of each mutant under different electron acceptor conditions. Growth studies revealed significant effects of these mutations with different electron acceptors compared to wild type MR-1, suggesting a complex network of electron transfer reactions. In agreement with previous observations, a key *c*-type cytochrome, *CymA*, showed decreased growth dynamics in five different terminal reductases except TMAO. In addition the *mtrDEF* gene cluster, which is similar to the metal reduction gene cluster (*mtrC/omcB*, *mtrAB*) were also partially defective in growth with Mn(IV) and Cr(IV). Mutants in the high affinity *cbb3* cytochrome oxidase components exhibit a defect both aerobically and anaerobically with TMAO and Cr(VI), suggesting a role for this complex in both suboxic and anaerobic respiratory processes. In addition, an insertional mutant was obtained for a periplasmic tetraheme flavocytochrome gene, *SO3056*. Genetic analyses indicate that this mutant is defective in Fe(III) reduction and growth with thiosulfate, nitrate, fumarate, DMSO but not TMAO and oxygen. The phenotypes of *SO3056* mutant are largely similar to what have been known for *CymA* mutant. Thus, *SO3056* might function downstream of cytoplasmic membrane protein *CymA* and likewise control multiple branches of electron transfer chain. It is clear from this overview of the current data available that many of these *c*-type cytochromes participate in respiratory metabolism, detoxification or sensing processes that have not yet been explored in

detail. Therefore, the availability of these *c*-type cytochrome mutants and of additional sequenced *Shewanella* strains provides an excellent resource for comparative physiology studies and will greatly facilitate our goal of characterizing respiratory networks in *Shewanella*.

105 ^{GTL}

A Phylogenetic Gibbs Sampler for High-Resolution Comparative Genomics Studies of Transcription Regulation

William A. Thompson,¹ Sean P. Conlan,² Thomas M. Smith,^{2,3} Lee A. Newberg,^{2,3} Lee Ann McCue^{2,4*} (leeann.mccue@pnl.gov), and **Charles E. Lawrence**¹

¹Center for Computational Molecular Biology and the Division of Applied Mathematics, Brown University, Providence, Rhode Island; ²The Wadsworth Center, New York State Department of Health, Albany, New York; ³Department of Computer Science, Rensselaer Polytechnic Institute, Troy, New York; and ⁴Pacific Northwest National Laboratory, Richland, Washington

Project Goals: 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 will further extend comparative genomics technologies that we have introduced over the last several years, developing and applying statistical approaches to analysis of correlated sequence data (i.e. sequences from closely related species). We also plan to combine functional genomic and proteomic data with sequence data from multiple species; combining these complementary data types promises to improve our ability to predict regulatory sites of small or genus-specific regulons.

High-throughput sequencing initiatives are enabling ambitious comparative genomics projects, such as the detection of functionally conserved regions in protein and DNA sequences. Of particular interest is the identification of transcription factor binding sites (TFBS) and *cis*-regulatory modules in the promoters of genes - a critical step for delineating the transcription regulatory network of an organism. While closely related species are most likely to share common transcription factors (and therefore common *cis*-regulatory elements), the recent speciation of closely related genomes results in correlation among the sequences which confounds the detection of functionally conserved motifs.

To facilitate high-resolution comparative genomics studies, we have developed a version of the Gibbs recursive sampler that incorporates phylogeny of the input sequences through the use of an evolutionary model, and calculates an ensemble centroid motif solution. This phylogenetic Gibbs sampler accepts aligned as well as unaligned orthologous sequence data; these may be orthologous sequences from a single gene or orthologous sequences for a group of co-regulated genes. The algorithm also requires a user-supplied tree describing the phylogenetic relationship of the orthologous sequences within each multiple-sequence alignment. For each alignment, the algorithm traverses this phylogenetic tree and calculates the joint probability of each nucleotide at each position, ultimately describing a motif as a product phylogeny model. We also employ an "ensemble centroid" motif estimator, i.e., the solution that is the set of sites that has the minimum total distance to the set of sites sampled from the posterior weighted ensemble of sites. The results below indicate that these two extensions provide significant additional power to the existing capabilities of the Gibbs recursive sampler.

We demonstrate the advanced features of the phylogenetic Gibbs sampler on the challenging problem of predicting motifs in orthologous data from a single gene. Using data that simulate this scenario, we show that false positive predictions, caused by correlation among the sequences, are dramatically reduced by the features described here. Specifically, we show that using a phylogenetic model and ensemble centroid solutions yields improved positive predictive values (PPV), an improved ability to avoid false positives, over a non-phylogenetic version of the Gibbs recursive sampler and a phylogenetic version of the Gibbs sampler that predicts maximum-likelihood alignments. We further demonstrate the ability of the phylogenetic Gibbs sampler to detect transcription factor binding sites in real orthologous sequence data from eight proteobacterial genomes, the majority of which are closely related enterobacteria.

106 MEWG

Challenges in Predictive Modeling for Engineering/Deciphering the Regulatory Networks

James C. Liao*

Chemical and Biomolecular Engineering, University of California, Los Angeles, California

Project Goals: Network analysis and prediction

The ultimate goal of systems biology and metabolic engineering is the prediction of cellular behavior, either for wild-type strains or the engineered strains. This ambitious task involves two levels: 1) prediction of regulatory networks, and 2) prediction of the behavior of the regulatory networks. We will discuss challenges in these problems and suggest some possible ways to tackle them. In particular, we will address how mathematical tools can interact with experimental data to maximize information output.

107 GTL

Molecular Mechanisms Regulating Gene Expression in *Geobacter sulfurreducens* under Environmentally Relevant Conditions

Toshiyuki Ueki^{1*} (tueki@microbio.umass.edu), Ching Leang,¹ Byoung-Chan Kim,¹ Richard Glaven,¹ Haiping Ke,¹ Katy Juárez,² and **Derek R. Lovley**¹ (dlovley@microbio.umass.edu)

¹Department of Microbiology, University of Massachusetts, Amherst, Massachusetts and ²Universidad Nacional Autónoma de México, Mexico

Project Goals: Our project goal is to predictively model how *Geobacter* species respond to natural environmental conditions or conditions that may be artificially imposed to promote in situ uranium bioremediation and electricity harvesting from waste organic matter and renewable biomass. Especially in this subproject, the molecular mechanisms for regulation of gene expression in *Geobacter sulfurreducens* are being investigated in order to better understand physiological functions of *Geobacter* species during in situ bioremediation and on the surface of electricity harvesting electrodes. These studies continually increase our knowledge of regulatory mechanisms

at the molecular level which can provide important insights into the physiology and ecology of *Geobacter* species in these environments.

The mechanisms for regulation of gene expression in *Geobacter sulfurreducens* are being investigated in order to better understand the response of *Geobacter* species to various environmental conditions that *Geobacter* species face during *in situ* uranium bioremediation and on the surface of electrodes harvesting electricity from waste organic matter and renewable biomass. Previous studies have demonstrated that these studies on regulatory mechanisms can provide important insights into the physiology and ecology of *Geobacter* species in these environments.

It is well known that sigma factors play an important role in regulating gene expression by recognizing promoter elements and initiating transcription. *G. sulfurreducens* has six genes encoding homologues of the sigma factors, RpoS, RpoN, RpoH, RpoD, RpoE, RpoN, and FliA. We have previously shown that RpoS is the stationary-phase sigma factor and is required for environmentally significant processes such as Fe(III) reduction and growth with oxygen. We are currently focusing on the alternative sigma factors, RpoN and RpoH. RpoN was constitutively expressed under all growth conditions examined. The *rpoN* gene appeared to be essential for growth, because a deletion mutant strain of the *rpoN* gene could not be obtained. When *rpoN* was overexpressed, growth was inhibited under a variety of environmentally relevant conditions, including growth in the absence of fixed nitrogen. These results suggest that proper expression of *rpoN* is important for optimum growth in the environment. A whole-genome DNA microarray comparison of gene expression between the wild-type strain and the *rpoN*-overexpressing strain revealed that RpoN regulated many genes involved in various cellular activities including pili and flagella biosyntheses. This is significant because of the important role of pili and flagella in growth in the subsurface and on the surface of electrodes.

Expression of the *rpoH* gene was induced, when the growth temperature was increased from 30°C to 42°C. A mutant in which *rpoH* was deleted grew normally at 30°C, but could not adapt to growth at 42°C. Furthermore, the expression of a number of heat-shock genes was undetectable or drastically decreased in the *rpoH*-deficient mutant. These results demonstrate that RpoH is the heat-shock sigma factor in *G. sulfurreducens* and essential for adaptation to growth at higher temperatures.

Two-component systems are an important strategy for adaptation to changes in environmental conditions. They typically consist of a sensor kinase, which senses an environmental signal, and a response regulator, which regulates gene expression to adapt to an environmental change. *G. sulfurreducens* contains an unusually large number of two-component systems, which may reflect the need to adapt to a myriad of different conditions in subsurface environments. A response regulator designated PilR was found to be important for the expression of *pilA*, which encodes the structural protein of the *G. sulfurreducens* pili. These pili are also referred to as microbial nanowires, because they are electrically conductive and are required for Fe(III) oxide reduction as well as optimal current production in microbial fuel cells. When *pilR* was deleted, it yielded a phenotype similar to the *pilA* mutant.

Fe(III) reduction as well as redox sensing were regulated by a two-component system consisting of a sensor kinase, which contains a unique sensor domain with ϵ -type heme binding motifs and a region with similarity to the “redox box” of RegB kinase, and a response regulator in the enhancer binding protein family. Another response regulator with a unique output domain was found to be involved in a variety of cellular processes including cell division, biofilm formation, and hydrogen-dependent growth. Furthermore, a sensor kinase, which microarray analysis indicated was up-regulated during nitrogen fixation, was necessary for optimum growth in the absence of fixed nitrogen. These results demonstrate the important role of the two-component systems in *G. sulfurreducens* for adapting to environmental changes.

G. sulfurreducens contains more than a hundred *c*-type cytochromes. Previous studies have demonstrated that deletion of some cytochrome genes inhibits Fe(III) reduction. However, it has recently become clear that some of these cytochromes are not directly involved in electron transfer, but rather are required for expression of other cytochromes that are key in electron transfer. Several more examples of this phenomenon have been discovered in the last year. For example, a mutant in which the gene for the cytochrome MacA was deleted could not express OmcB, an outer-membrane cytochrome that is required for Fe(III) reduction. When *omcB* was expressed with an exogenous constitutive promoter, the *macA*-deficient mutant reduced Fe(III) as well as the wild-type. A mutant in which the gene for the outer-membrane cytochrome OmcF was deleted was defective in Fe(III) reduction and electricity production. Microarray analysis revealed that genes involved in Fe(III) reduction and electricity production were down-regulated in the OmcF mutant.

Previous analysis of *in situ* gene transcript levels during *in situ* uranium bioremediation has demonstrated that levels of transcripts for the citrate synthase gene can be used to monitor the rates of *Geobacter* metabolism in the subsurface. Deletion of the citrate synthase gene eliminated the capacity for *G. sulfurreducens* to grow not only on acetate, but also with hydrogen as the electron donor when Fe(III) was the electron acceptor. Expression of the citrate synthase gene was regulated mainly at the level of transcription by a repressor. The binding site for the repressor was determined in the promoter of the citrate synthase gene and genome sequence analyses further identified sequences similar to the binding site within the promoter regions of other genes, most of which are likely involved in metabolic pathways including the TCA cycle, in *G. sulfurreducens* as well as other *Geobacter* species.

Overall, these studies have significantly added to the understanding of key modes of regulation in *Geobacter* species that are necessary for important physiological functions in the subsurface and on the surface of energy-harvesting electrodes. This permits us to continually increase the sophistication of our models for predicting how *Geobacter* species will respond to different natural environmental conditions or conditions that may be artificially imposed to promote bioremediation or energy harvesting.

108 ^{GTL}

Computational Analysis of Transcription Regulation of *Geobacter sulfurreducens*

Julia Krushkal^{1*} (jkrushka@utmem.edu), Marko Puljic,¹ Ronald M. Adkins,¹ Jeanette Peeples,¹ Bin Yan,^{1,2} Ching Leang,³ Laurie N. DiDonato,³ Cinthia E. Núñez,^{3,4} Toshiyuki Ueki,³ Radhakrishnan Mahadevan,⁵ Brad Postier,³ Barbara Methé,⁶ and **Derek R. Lovley**³ (dlovley@microbio.umass.edu)

¹University of Tennessee Health Science Center, Memphis, Tennessee; ²NIDCD, National Institutes of Health, Bethesda, Maryland; ³University of Massachusetts, Amherst, Massachusetts; ⁴Instituto de Biotecnología/UNAM, Cuernavaca, México; ⁵University of Toronto, Toronto, Canada; and ⁶The Institute for Genomic Research, Rockville, Maryland

Project Goals: The overall purpose of this project is to develop experimental and computational tools to predictively model the behavior of complex microbial communities involved in microbial processes of interest to the Department of Energy. The five year goal is to deliver *in silico* models that can predict the behavior of two microbial communities of direct relevance to Department of Energy interests: 1) the microbial community responsible for *in situ* bioremediation of uranium in contaminated subsurface environments; and 2) the microbial community capable of harvesting electricity from waste organic matter and renewable biomass. The research in this abstract

summarizes research under Subproject IV. The purpose of this subtask is to use computational methods to better understand transcriptional regulation of the expression of environmentally relevant genes in *Geobacter* species.

Geobacter species are of interest because of their role in *in situ* bioremediation of uranium and harvesting electricity from waste organic matter and renewable biomass. As part of our involvement in the Genomics:GTL *Geobacter* Project, we are investigating networks of regulatory interactions in this versatile group of microorganisms in order to elucidate molecular mechanisms of their regulatory response to environmental changes.



Figure 1. An example of GSEL query output listing predicted regulatory sites and operon structure in a region of the *G. sulfurreducens* genome.

To unveil transcriptional regulatory interactions affecting *Geobacter sulfurreducens* gene expression, we are employing a variety of computational strategies and utilizing a vast array of genome sequence data and gene expression information obtained in this project. In order to understand the complex interplay of multiple regulatory mechanisms, we not only catalogue individual genome sequence elements predicted to be involved in regulatory processes, but also apply data mining tools to those elements that appear to be involved in multiple regulatory pathways. This approach also allows us to identify those target genes that may be involved in important regulatory response mechanisms in a variety of conditions. In order to address these important questions, we have developed a database and an accompanying online query system, GSEL (*Geobacter* Sequence Elements) that compile information on putative transcription regulatory elements in the genome of *G. sulfurreducens* predicted by 10 different computational approaches based on pure *in silico* predictions and analysis of empirical data.

This online system allows users to query the genome of *G. sulfurreducens* using a specified genome region, operon number, or gene identifier (GSU). The output provides the predicted operon organization and the list of regulatory elements in the respective genome region ranked by the number of methods that predicted the site.

Individual regulatory sequence elements are predicted using analyses of individual microarray data sets or sequence data alone. For example, *G. sulfurreducens* has molybdate-responsive transcription factor, ModE (GSU2964). We identified 80 likely ModE binding sites in the genome of *G. sulfurreducens*, including likely functional sites in the upstream regions of (1) the *modABC* operon; (2) an operon containing gene NP_954455, a distant homolog of the *moaA* gene that encodes the molybdenum cofactor biosynthesis protein A in *Archaeoglobus fulgidus* and *Sulfolobus tokodaii*; (3) an

operon that encodes a putative membrane protein (NP_954447) with homology to permeases of the drug/metabolite transporter (DMT) family (COG0697); and (4) glycine (CCC) tRNA.

Similarly, *G. sulfurreducens* contains RpoS (σ S), a global regulator of gene expression in *G. sulfurreducens*, in addition to major housekeeping sigma factor RpoD (σ 70). Our analysis of conservation and divergence of possible functional RpoD amino acid residues suggested that most of them are substantially conserved between *E. coli* and *G. sulfurreducens*. However, some degree of sequence variation between the two species was observed in several amino acid sites of RpoS proteins that might be important for promoter recognition. In addition, the similarity or identity of a number of residues between *G. sulfurreducens* RpoS and RpoD that might be functionally important suggests that the differences between the promoters recognized by RpoS and RpoD in *G. sulfurreducens* may be subtle. Using microarray gene expression information, we have been able to suggest a number of RpoS-regulated promoter elements as well as elements regulated by other sigma factors including RpoD. Experimental analysis of several promoters predicted to be RpoS-regulated and those predicted to be RpoD-regulated fully validated computational predictions.

In another example of analysis, we predicted gene regulatory interactions using information from the RelGsu regulon. RelGsu is the single *Geobacter sulfurreducens* homolog of RelA and SpoT proteins found in other organisms to be involved in regulation of levels of guanosine 3', 5' bispyrophosphate, ppGpp, a molecule that signals slow growth and stress response under nutrient limitation in bacteria. We used information obtained from genome wide expression profiling of the *rel_{Gsu}* deletion mutant to identify putative regulatory sites involved in transcription networks modulated by RelGsu or ppGpp. We identified likely sites regulated by Fur (ferric uptake repressor) in the upstream regions of upregulated operons and RpoS-regulated promoters in the upstream regions of the downregulated operons of the *rel_{Gsu}* deletion mutant. These findings suggest that Fur- and RpoS-dependent gene expression in *G. sulfurreducens* is affected by ppGpp-mediated signaling.

Among multiple other sequence and gene expression analyses by our group of regulatory interactions influenced by specific transcription factors, our most recent analyses involve prediction of the promoter elements regulated by alternative sigma factor RpoN (σ 54). We predicted 467 RpoN-regulated promoter elements that had the same orientation with their downstream target genes or operons, including 110 such elements in the noncoding regions. We identified those promoters for which the expression of their target genes was significantly altered in the RpoN gene overexpression microarrays. Further analyses focus on the function of the specific genes whose regulation may be significantly affected by RpoN and on their possible role in different environmental conditions.

Data mining of predicted regulatory interactions allowed us to identify genome regulatory regions and their target operons that are involved in a variety of regulatory pathways. This powerful approach allows us to identify gene products that may be central to *G. sulfurreducens* response to a variety of trigger conditions, to find genes and operons whose expression may be altered in response to very specific sets of conditions, and to suggest the molecular mechanisms of their regulation.

109 ^{GTL}

Identification of Small Non-Coding RNAs and Acceptance Rate Studies in Members of the *Geobacteraceae*

Barbara Methé^{1*} (bmethe@tigr.org), Robert DeBoy,¹ Sean Daugherty,¹ Ty Arrington,¹ Kelly Nevin,² Jonathan Badger,¹ and **Derek Lovley**²

¹The Institute for Genomic Research, Rockville, Maryland and ²University of Massachusetts, Amherst, Massachusetts

Project Goals: This abstract submission is a part of the project “Genome-Based Models to Optimize In Situ Bioremediation of Uranium and Harvesting Electrical Energy from Waste Organic Matter” (PI- Lovley) and is focused on the application of comparative genomic and functional genomic techniques to improve our understanding of the evolution and regulation of members of the *Geobacteraceae*.

Members of the *Geobacteraceae* are the subject of intense study as they are the dominant dissimilatory metal-reducing microorganisms in subsurface environments in which organic contaminants are being degraded with the reduction of Fe(III) and in aquatic sediments where dissimilatory metal reduction is important. Further they are of great interest for the practical roles that they can play as agents of bioremediation and in energy production. The completion of genome sequence from multiple members of the *Geobacteraceae* provides the opportunity to apply genome-level analyses to obtain fundamentally new insights into their evolution and regulation.

Predicting the presence and function of protein coding genes has traditionally been an important area of focus in microbial genomics. More recently, a significant interest has developed in the study of chromosomally located small non-coding (sRNAs) which is being aided by the completion of genome sequence from related organisms. Increasing evidence suggests that sRNAs exist in numerous organisms where they play important regulatory roles including responses and adaptations to different stresses.

To identify novel sRNAs in *Geobacter sulfurreducens* and other *Geobacteraceae* we are adapting the computer program, sRNAPredict2, which uses an integrative computational approach to identify sRNAs in bacterial genomes. This program relies on sequence conservation outside of protein-coding genes and the locations of predicted Rho-independent terminators. Several supporting programs which provide input files for sRNAPredict2 are also being used to improve sRNA prediction. These supporting programs provide the locations of protein-encoding genes, rRNAs and tRNAs, a small number of previously predicted sRNAs, predicted terminator sequences and putative conservation of RNA secondary structure in intergenic regions. Currently a total of 80 putative sRNA genes have been identified in *G. sulfurreducens*, of which more than 30 were predicted in both sets of pairwise comparisons with the genomes of *G. metallireducens* and *G. uraniumreducens* and the sequences of six of these are closely related to one another based on primary and predicted secondary sequence structure. The longer term goal of this effort is to experimentally characterize these predictions using a combination of traditional molecular biological techniques such as Northern blots and primer extensions and to use microarray technology in the form of an oligonucleotide tiling array of the complete *G. sulfurreducens* genome.

Microbial genomes are not static entities. The rate and degree of genomic plasticity can best be understood in the framework of evolutionary processes. Rates of nucleotide substitution in genome sequence are a composite of the occurrence of mutations, random genetic drift of neutral or nearly

neutral alleles and purifying selection against deleterious alleles. In a small number of cases, directional selection (positive selection) occurs when natural selection favors a single allele and therefore allele frequency continuously shifts in that direction. Nonsynonymous mutations result in amino acid replacement while synonymous or silent mutations cause no change in the specified amino acid. Since advantageous mutations undergo fixation in a population more rapidly than neutral mutations the rate of nonsynonymous substitution will exceed that of synonymous substitution if advantageous selection plays a role in the evolution of the protein in question. One way to detect positive Darwinian selection is to determine if the number of substitutions per nonsynonymous site is significantly greater than the number of substitutions per synonymous site.

We are studying the rates of nonsynonymous to synonymous (d_N/d_S) substitutions) across gene families of interest in members of the *Geobacteraceae* and other selected genomes. In a study examining d_N/d_S ratios in c-type cytochromes for example, 26 orthologous gene clusters were identified in at least 5 of the following 10 genomes: *Geobacter sulfurreducens*, *G. sp. FRC-32*, *G. metallireducens*, *G. uraniumreducens*, *Pelobacter carbinolicus*, *P. propionicus*, *Desulfuromonas acetoxidans*, *Desulfovibrio desulfuricans*, *D. vulgaris* and *Rhodospirillum rubrum*. All 26 clusters were determined to have d_N/d_S ratios of less than one signifying that members of these clusters are undergoing purifying selection. However, examination of amino acid sites within the protein sequences revealed sites with statistically significant evidence of positive selection. For example, four sites have been determined to be undergoing positive selection within a c-type cytochrome for which a three-dimensional structure has been solved (10S6). These sites are located at positions 22, 25, 28 (near ligand to Heme III at His20) and 45 (near Heme IV at His47). A further systematic evaluation of sequences from the *Geobacteraceae* is currently underway. These investigations are of relevance to protein engineering as understanding which residues have changed historically may suggest sites that can be mutated to change the specificity and metabolic characteristics of the protein in question.

110 ^{GT}L

Bacterial Cell Cycle Control System and a Control System Simulation Model

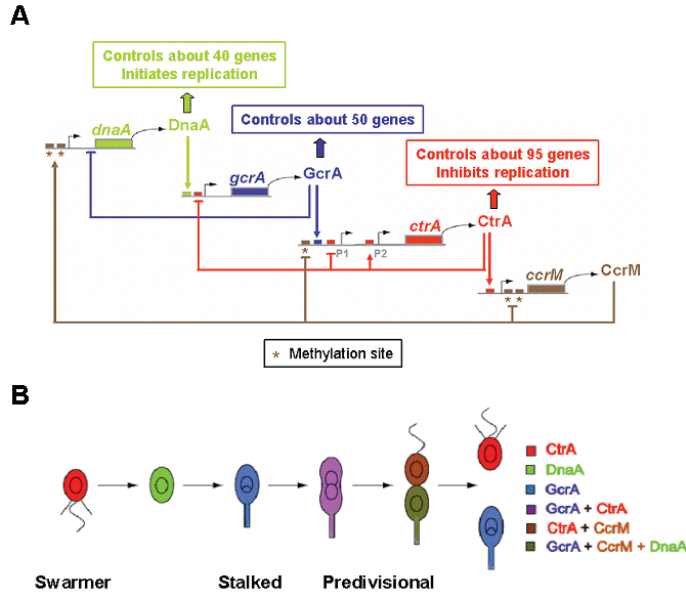
J. Collier,¹ X. Shen,² M. Horowitz,² L. Shapiro,¹ and H. H. McAdams^{1*} (hmcadams@stanford.edu)

¹Department of Developmental Biology and ²Department of Electrical Engineering, Stanford University, Stanford, California

Project Goals: Identification of the overall regulatory and metabolic networks in *Caulobacter crescentus*, largely through gene expression microarray assays and bioinformatic analysis

A core cell cycle engine involving interlinked master regulator proteins drives cell cycle progression in bacterial cells as in eukaryotic cells. The cell cycle control system in the model bacteria *Caulobacter crescentus* is hierarchically organized, and it activates functional genetic modules ‘just in time’ when needed, including two processive subsystems, chromosome replication and cytokinesis. The core engine and modular cell cycle subfunctions are repeatedly re-synchronized by non-genetic mechanisms, and checkpoint signaling between modules also acts to assure proper ordering of subfunctions. Each top-level regulatory protein is regulated at multiple levels, e.g., transcription, stability, and activity. In *Caulobacter*, three essential oscillating transcriptional regulators, DnaA, GcrA, and CtrA, control over 200 temporally controlled genes. CtrA also blocks the initiation of chromosome replication, while DnaA promotes it, by directly binding to the origin of replication (*Cori*) and allowing

replisome formation. Elimination of DnaA halts the cell cycle, and control of DnaA stability is a mechanism used by stress sensors to halt the cell cycle. DnaA activates the synthesis of GcrA, which in turn activates the synthesis of CtrA that activates the synthesis of the CcrM DNA methyltransferase. This cascade of master regulator proteins drives the forward progression of the *Caulobacter* cell cycle as shown in Figure 1.



Cyclical oscillator drives the cell cycle. We have shown that a critical element of the *Caulobacter* cell cycle regulatory circuit is the link between *dnaA* expression and the progression of DNA replication. We have demonstrated that the synthesis of DnaA is coordinated with its function by an epigenetic mechanism of regulation, through the methylation of the *dnaA* promoter by the cell cycle-regulated CcrM protein. The *dnaA* promoter contains two DNA methylation sites. We have also shown that these two methylation sites are essential for efficient *dnaA* transcription. When the two meth-

ylation sites in the *dnaA* promoter are removed by targeted mutagenesis at the *dnaA* locus, the level of DnaA becomes limiting in cells, resulting in abnormal cell phenotypes. Furthermore, the transcription of *dnaA*, which resides near the *Cori*, is efficient when the *dnaA* promoter is in the fully-methylated state prior to the initiation of DNA replication, but inefficient when the *dnaA* promoter becomes hemimethylated upon passage of the replication fork, soon after DNA replication initiation. Accordingly, the chromosomal location of *dnaA* near the *Cori* is an important component of this regulatory mechanism, which contributes to the changes in DnaA cellular content during the cell cycle. We conclude that the induction of the master regulatory cascade is tied to the replication status of the chromosome by CcrM, which is the last element of this cascade. One major consequence of this finding is that the DnaA/GcrA/CtrA/CcrM master regulatory cascade is a cyclic process, where CcrM activates DnaA at the end of the cell cycle to start a new cell cycle (Figure 1).

Engineering simulation of the *Caulobacter* cell cycle control system. We have approached analysis of the circuitry from an engineering perspective, using computer tools and analysis paradigms drawn from electrical engineering. The results demonstrate both how the *Caulobacter* cell cycle regulatory mechanisms functions as a system and how the circuit is designed for reliable or robust operation in spite of noisy components and highly uncertain operating environments. From the EE circuit design perspective, the cell cycle control circuit approximates a fundamental mode asynchronous state machine with design features long known to electrical engineers as necessary for robust operation. Engineers exploit this FASM class of circuits in situations that are parallel to the cell control problem. For example, there are unclocked electrical systems with concurrent modular subsystems and wide variability in component performance as in the biological cell cycle circuitry. The *Caulobacter* cell cycle circuit design is robust to parameter variation, and it can provide reliable operation over a wide range of growth rates (i.e., generation times) consistent with the requirements of an organism adapted to

low and uncertain nutrient levels. Additional reliability is provided by the elaborate and redundant regulatory mechanisms that control the most important regulatory proteins, such as CtrA and DnaA.

Since cell cycle regulation involves decisively turning modular functions on and off, the regulatory mechanisms tend to function as binary switches as in many areas of developmental regulation in all organisms. Rapid switching between qualitatively different stable cell states can result from bistable regulatory circuits, that is, circuits that exhibit hysteresis and require changes in input signals to initiate a transition between states. A central element of the design of the *Caulobacter* cell cycle regulatory circuit is a bistable switching element resulting from positive autoregulation of the *ctrA* gene encoding the CtrA master regulator protein. Changes in a phosphosignal originating from the CckA histidine kinase initiate switching between high and low CtrA activity.

Since *Caulobacter* cells divide asymmetrically, producing siblings with significantly different morphology and cell fates, its cell cycle control circuitry has to initialize each daughter cell's control system differently consistent with each cell's individual regulatory program. Dynamic localization of regulatory proteins and proteolytic subsystems to the cell poles is essential to asymmetric cell division. The distinctive identity of the subsequent daughter cells, each containing one of the chromosomes of the predivisional cell, begins at the instant of cytoplasmic compartmentalization about 18 minutes before daughter cell separation. Immediately upon compartmentalization, differentiation begins owing to isolation of key phosphorylation-dependent regulatory proteins from their cognate kinases or perhaps to differential sequestering of a phosphatase. Large differences in binding affinity between the phosphorylated and unphosphorylated response regulators causes gene expression profiles in the compartments to diverge yielding different development programs thereafter with profound consequences for the fates of the two daughter cells.

We have developed a system-level computer simulation of the *Caulobacter* cell cycle and asymmetric cell division control system using the Matlab control system simulation tools, Simulink and Stateflow, that are widely used in engineering analysis. The simulation model provides verification of the operation of the cell cycle control system design by comparing a molecular level simulation of the system with experimental observations of changing protein and mRNA levels over the cell cycle. We find excellent correlation between protein and mRNA levels predicted by the simulation model and experimental results. The simulation model of the *Caulobacter* cell cycle is inherently extensible. In the current version, we include detailed models of the chromosome replication and cytokinesis submodules because progress of these two subsystems is so essentially integrated with progress of the cell cycle engine, but future extensions can add regulation of phosphosignaling, polar organelle development, and blocking of cell cycle progress under limited resource and stress conditions. This top-down incremental development paradigm is a promising avenue for development of whole cell behavioral simulation models that can both emulate the observed behavior of the cell and predict the outcome of genetic changes.

Modeling biological regulatory circuits entirely as systems of ordinary differential equations is inherently self-limiting. Biological control circuits have elements that are more readily simulated with hybrid models that combine differential equation-based models of subsystems with state machine models and *ad hoc* behavioral models. Our results suggest that the attempts to model regulatory networks assuming a transcriptional regulation paradigm without consideration of both dynamic localization of regulatory proteins and of epigenetic chromosomal modifications are unlikely to approximate biological reality.

111 ^{GTL}**Automated Accurate, Concise and Consistent Product Description Assignment for Microbial Regulatory Proteins**

Loren Hauser* (hauserlj@ornl.gov), Frank Larimer, and **Miriam Land**

Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee

Project Goals: The project is to develop and continually improve an automated microbial genome annotation pipeline. The current project is the newest bioinformatic tool to be added to the pipeline. This tool is designed to identify, categorize and assign consistent product descriptions for most of the regulatory proteins in every annotated microbial genome.

US DOE's Joint Genome Institute (JGI), which consists of groups at the Production Genome Facility in Walnut Creek CA, LBL, LLNL, LANL, ORNL and Stanford, is scheduled to sequence over 120 microbial genomes in 2007. ORNL has developed and implemented an automated annotation pipeline for the initial annotation of these genomes. The pipeline's multiple tools and database queries are used to create a reference web site for each genome that is used as the basis for both automated and manual annotation. Although most consider manual annotation optimal, it varies in quality, is time consuming, and does not scale with the increase in genomic sequencing. Therefore, it is not feasible for the majority of genomes. However, accurate automated annotation of most of an organism's gene products which contain characterized protein domains is feasible. Complex multidomain proteins, such as signal transduction histidine kinases and other regulatory proteins represent a unique challenge due to the number and variety of domains that are possible in a single protein. A rule based annotation system combining the output from searches of Interpro, COGs, and Swissprot-TREMBL, as well as, TMHMM predictions and the domain architecture of individual proteins has been developed in order to provide accurate, concise and consistent product descriptions for most microbial regulatory proteins. This multidimensional approach increases both the accuracy and sensitivity when compared to the use of single systems such as SMART or simple blast searches. The system provides a list of all the regulatory proteins it has identified, the product descriptions that will be assigned, and all the information used in their identification. The product description nomenclature scheme is being applied to all genomes in IMG so that searches will find all members of gene families regardless of the quality or age of the original annotation. In addition, it will provide a numerical synopsis in the form of a table of the regulatory proteins organized by the effector or output domain, which can be directly used in genome publications. This system has recently been incorporated into the automated annotation pipeline. Similar annotation tools are under development here and at other JGI sites.

GTL Milestone 3

Develop the Knowledgebase, Computational Methods, and Capabilities to Advance Understanding and Prediction of Complex Biological Systems

Section 1

Computing Infrastructure, Bioinformatics, and Data Management

112 ^{GTL}

Center for Computational Biology at the University of California, Merced

Michael Colvin^{1*} (mcolvin@ucmerced.edu), Arnold Kim,¹ Masa Watanabe,^{1*} and Felice Lightstone²

¹School of Natural Sciences, University of California, Merced, California and ²Biosciences Directorate, Lawrence Livermore National Laboratory, Livermore, California

Project Goals: The goals of the UCM-CCB are 1) to help train a new generation of biologists who bridge the gap between the computational and life sciences and to implement a new biology curriculum that can both influence and be adopted by other universities, 2) and to facilitate the development of multidisciplinary research programs in computational and mathematical biology.

The Center for Computational Biology (UCM-CCB) was established at the newest campus of the University of California in fall, 2004. The UCM-CCB is sponsoring multidisciplinary scientific projects in which biological understanding is guided by mathematical and computational modeling. The center is also facilitating the development and dissemination of undergraduate and graduate course materials based on the latest research in computational biology. This project is a multi-institutional collaboration including the new University of California campus at Merced, Rice University, Rensselaer Polytechnic Institute, and Lawrence Livermore National Laboratory, as well as individual collaborators at other sites.

The UCM-CCB is sponsoring a number of research projects that emphasize the role of predictive simulations in guiding biological understanding. This research is being performed by post-docs, graduate and undergraduate students and includes mathematical models of cell fate decisions, molecular models of multiprotein machines such as the nuclear pore complex, new mathematical methods for simulating biological processes with incomplete information, and mathematical approaches for simulating the interaction of light with biological materials. The UCM-CCB has run workshops to facilitate computational collaborations with many of the experimental biology programs at UC Merced and is hosting an ongoing seminar series that has brought many prominent computational biologists to speak at UC Merced over the past year.

Additionally, the UCM-CCB is having a central role in enabling the highly mathematical and computationally intensive Biological Science major, which is currently the largest major at UC Merced. Over the past year, members of the UCM-CCB have visited several other universities to describe this new major. All electronic, modular course materials produced by the UCM-CCB being released under an open public license and have been facilitating linkages to feeder schools at the state university, community college, and high school levels. Last summer the UCM-CCB ran a six-week compu-

tational biology internship program, that culminated in a one-week visit to the National Center for Supercomputing Applications at the University of Illinois.

The long-term goal of the UCM-CCB is to help train a new generation of biologists who bridge the gap between the computational and life sciences and to implement a new biology curriculum that can both influence and be adopted by other universities. Such scientists will be critical to the success of new approaches to biology, exemplified by the DOE Genomics:GTL program in which comprehensive datasets will be assembled with the goal of enabling predictive modeling of the behavior of microbes and microbial communities, as well as the biochemical components of life, such as multiprotein machines.

113 ^{GTL}

Projects from the DOE-BACTER Institute at the University of Wisconsin, Madison

Julie C. Mitchell* (mitchell@math.wisc.edu), Julie Simons, Paul Milewski, Peter Koenig, and Qiang Cui

University of Wisconsin, Madison, Wisconsin

Project Goals: BACTER was initiated through a DOE call for proposals to address the critical need for computational modelers in the area of bioenergy systems modeling. BACTER faculty, students and postdocs work on DOE-relevant modeling projects, ranging from molecules, to pathways and cells, to entire populations.

Selected here are two of many projects ongoing at the DOE-BACTER Institute. BACTER was initiated through a DOE call for proposals to address the critical need for computational modelers in the area of bioenergy systems modeling. BACTER faculty, students and postdocs work on DOE-relevant modeling projects, ranging from molecules, to pathways and cells, to entire populations.

***R. sphaeroides* and *E. coli*: a comparison of population-level behavior**

Understanding the behavior of bacteria in populations is of importance in biofilm formation, our potential to harness bacteria for use in such contexts as bioremediation, and understanding how simple organisms display complex behavior. We are interested in modeling the chemotactic behavior of *Rhodobacter sphaeroides* in an effort to elucidate why this species behaves differently from the better understood *Escherichia coli* in different environments. Specifically we performed “swarm-plate” experiments, inoculating *R. sphaeroides* and *E. coli* in agar plates with varying concentrations of the chemoattractant L-aspartate. Separately we also measured growth rates for the bacteria in liquid cultures.

From the data obtained, we examine the behavior seen via a set of partial differential equations known as the Keller-Segel equations. These equations allow us to differentiate growth effects from chemotaxis and point to sensitivities of *R. sphaeroides* as compared to *E. coli* in terms of their abilities to create and respond to gradients of chemoattractant and move within the agar media.

Perturbative analysis of potential of mean force simulations

Enzymes reach their high efficiency in catalyzing reactions. Bioenergetic enzymes interconvert energy with unsurpassed efficiency. Gaining insight into how enzymes modulate the energetics of a reaction is a key step to their understanding.

Molecular dynamics simulations of chemical reactivity have become routine. The simulation of the free energy along a reaction coordinate, not only describes the chemical reaction *per se*, but is also able to capture the dynamic response of the reacting species and their environment. Despite the advantages of these simulations, the methods for their analysis lag behind. For instance, there is no established methodology for probing the quality of the potential functional or the importance of interactions and contributions.

We suggest a perturbative approach for examining the importance of interactions for the energetics of a reaction. Beyond simple *in silico* mutations, this technique allows the exploration of components not easily accessible to experimental study, e.g. the electrostatic effect of a whole part of a protein or solvation.

We present the application of this method for studying reactions in a number of enzymes. We have also applied this methodology to study the mechanism of proton blockage in aquaporin, which is an ongoing subject of discussion in the literature. This membrane channel efficiently conducts water molecules but prevents the decoupling of bioenergetic processes due to proton leakage. With the perturbative approach introduced here, the different contributions to the energetics could be dissected.

114 ^{GTL}

VIMSS Computational Core

Paramvir S. Dehal^{1,2*} (PSDehal@lbl.gov), Eric J. Alm^{1,3} Dylan Chivian^{1,2} Katherine H. Huang^{1,2} Y. Wayne Huang^{1,2} Janet Jacobsen^{1,4} Marcin P. Joachimiak^{1,2} Keith Keller^{1,4} Morgan N. Price^{1,2} and **Adam P. Arkin**^{1,2,4,5,6} (aparkin@lbl.gov)

¹Virtual Institute for Microbial Stress and Survival, <http://vimss.lbl.gov/>; ²Lawrence Berkeley National Laboratory, Berkeley, California; ³Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts; ⁴University of California, Berkeley, California; ⁵Howard Hughes Medical Institute, Chevy Chase, Maryland; and ⁶Department of Bioengineering, University of California, Berkeley, California

Project Goals: Environmental Stress Pathway Project (ESPP) is developing computational models that describe and predict the behavior of gene regulatory networks in microbes in response to the environmental conditions found in DOE waste sites. The research takes place within the Virtual Institute for Microbial Stress and Survival (VIMSS). Based at Lawrence Berkeley National Laboratory (LBNL), VIMSS supports an integrated and multi-institutional program to understand the ability of bacteria and other microorganisms to respond to and survive external stresses. VIMSS was established in 2002 with funding from the U.S. Department of Energy Genomics: GTL Program for Rapid Deduction of Stress Response Pathways in Metal/Radionuclide Reducing Bacteria. LBNL is operated by the University of California for the U.S. Department of Energy.

Background: The VIMSS Computational Core group is responsible for data management, data integration, data analysis, and comparative and evolutionary genomic analysis of the data for the VIMSS project. We have expanded and extended our existing tools sets for comparative and evolutionary genomics and microarray analysis as well as creating new tools for our proteomic and metabolomic data sets. Our analysis has been incorporated into our comparative genomics website MicrobesOnline (<http://www.microbesonline.org>) and made available to the wider research community. By taking advantage of the diverse functional and comparative datasets, we have been able to pursue large evolutionary studies.

Data Analysis: During the course of analysis of various stress responses of DvH the computational core has continued to develop new statistical analyses of data that take advantage of the predicted regulatory structures (operons, regulons, etc.) from our comparative analyses. This year we have used these analyses to investigate the response of DvH to oxygen stress and chromium stress. Our analysis has focused on the combined results from both transcriptomic and proteomic datasets to interpret oxygen stress. Additionally, we have begun preliminary work to examine metabolomic datasets within the framework of predicted metabolic activities.

Data Management: All data generated by ESPP continues to be stored in our Experimental Information and Data Repository (<http://vimss.lbl.gov/EIDR/>). Researchers have access to datasets from biomass production, growth curves, image data, mass spec data, phenotype microarray data and transcriptomic, proteomic and metabolomic data. New functionality has been added for storage of information relating to mutants and protein complex data, in addition to new visualization for assessing existing data sets such as the phenotype microarrays.

The MicrobesOnline Database: The MicrobesOnline database (<http://www.microbesonline.org>) currently holds 330 microbial genomes and will soon be expanded to over 600 genomes, providing an important comparative genomics resource to the community. New functionality added this year includes the addition of a phylogenetic tree based genome browser that allows users to view their genes and genomes of interest within an evolutionary framework, tools allowing users to search for novel regulatory binding site motifs or matches to existing regulatory binding motifs across a user selected set of genes using our gene carts, tools to compare multiple microarray expression data across genes and genomes and more integration with the RegTransDb of experimentally verified regulatory binding sites.

MicrobesOnline continues to provide an interface for genome annotation, which like all the tools reported here, is freely available to the scientific community. To keep up with the rapidly expanding set of sequenced genomes, we have begun to investigate methods for accelerating our annotation pipeline. In particular we are researching methods to speed up the most time consuming process, homology searching through HMM alignments and all against all BLAST. Over the next year we will be releasing methods that will allow us to deal with the many millions of gene sequences generated from metagenomics.

Over the next year, several new features will be added to the MicrobesOnline resource. Microarray expression data will be added from the NCBI GEO database, in addition to datasets generated from the VIMSS team. To supplement the analysis tools we already have, enrichment of functional genes and operon-wise analysis, we will provide tools for comparing multiple experiments across multiple genomes. We will also expand our regulatory binding motif search to incorporate co-expression data to support predictions.

Evolutionary Analysis: The computational core continues work on understanding the evolution of regulatory networks. Transcription factors form large paralogous families and have complex evolutionary histories. Our analysis shows that putative orthology derived from bidirectional best hits across distantly related bacteria are usually not true evolutionary orthologs. Additionally, these false orthologs usually respond to different signals and regulate distinct pathways. Even in more closely related genomes, such as *E. coli* and *Shewanella oneidensis*, bidirectional best hits have a high error rate. By studying transcription factors with phylogenetic trees, we show that through the use of gene-regulon correlations, together with sequence analysis of promoter regions for confirmation, bacterial regulatory networks may evolve more rapidly than previously thought.

RegTransBase – A Resource for Studying Regulatory Interactions and Regulon Predictions in Bacteria

Michael J. Cipriano,^{1,5*} Alexei E. Kazakov,² Dmitry Ravcheev,² **Adam P. Arkin**^{1,3,4,5} (aparkin@lbl.gov), Mikhail S. Gelfand,² and Inna Dubchak^{1,5,6}

¹Lawrence Berkeley National Laboratory, Berkeley, California; ²Institute for Information Transmission Problems, Moscow, Russia; ³Howard Hughes Medical Institute, Chevy Chase, Maryland;

⁴Department of Bioengineering, University of California, Berkeley, California; ⁵Virtual Institute of Microbial Stress and Survival, <http://vimss.lbl.gov>; and ⁶Department of Energy Joint Genome Institute, Walnut Creek, California

Project Goals: Environmental Stress Pathway Project (ESPP) is developing computational models that describe and predict the behavior of gene regulatory networks in microbes in response to the environmental conditions found in DOE waste sites. The research takes place within the Virtual Institute for Microbial Stress and Survival (VIMSS). Based at Lawrence Berkeley National Laboratory (LBNL), VIMSS supports an integrated and multi-institutional program to understand the ability of bacteria and other microorganisms to respond to and survive external stresses. VIMSS was established in 2002 with funding from the U.S. Department of Energy Genomics: GTL Program for Rapid Deduction of Stress Response Pathways in Metal/Radionuclide Reducing Bacteria. LBNL is operated by the University of California for the U.S. Department of Energy.

RegTransBase, a database describing regulatory interactions in prokaryotes, is manually curated and based on published scientific literature. RegTransBase was created to supplement the data available in MicrobesOnline by including published experimental information on regulation. It describes a large number of regulatory interactions and contains experimental data from ~3000 articles from a wide range of taxa (including environmentally interesting organisms) which investigates regulation with known elements. RegTransBase additionally provides an expertly curated library of alignments of known transcription factor binding sites, and includes the exact location of the binding site on a published genome, the transcription factor, and links to published articles. RegTransBase builds upon these alignments by containing a set of computational modules for the comparative analysis of regulons among related organisms which guide a user through the appropriate steps of transferring known or high confidence regulatory binding site results to many other microbial organisms. An intuitive, interactive user-friendly interface makes this knowledge freely accessible to the larger microbiological research community.

RegTransBase creates access to the highest quality regulatory information about sequenced microbes and, with MicrobesOnline, is a critical tool to the inference of regulatory and stress response networks that are central goals of the VIMSS::ESPP project. RegTransBase is available at <http://reg-transbase.lbl.gov>.

116 ^{GTL}**MicrobesOnline: An Integrated Portal for Comparative Functional Genomics**

Marcin P. Joachimiak,^{1,2*} Katherine H. Huang,^{1,2} Eric J. Alm,^{1,3} Dylan Chivian,^{1,2} Paramvir S. Dehal,^{1,2} Y. Wayne Huang,^{1,2} Janet Jacobsen,^{1,4} Keith Keller,^{1,4} Morgan N. Price,^{1,2} and **Adam P. Arkin**^{1,2,4,5,6} (aparkin@lbl.gov)

¹Virtual Institute for Microbial Stress and Survival, <http://vimss.lbl.gov/>; ²Lawrence Berkeley National Laboratory, Berkeley, California; ³Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts; ⁴University of California, Berkeley, California; ⁵Howard Hughes Medical Institute, Chevy Chase, Maryland; and ⁶Department of Bioengineering, University of California, Berkeley, California

Project Goals: Environmental Stress Pathway Project (ESPP) is developing computational models that describe and predict the behavior of gene regulatory networks in microbes in response to the environmental conditions found in DOE waste sites. The research takes place within the Virtual Institute for Microbial Stress and Survival (VIMSS). Based at Lawrence Berkeley National Laboratory (LBNL), VIMSS supports an integrated and multi-institutional program to understand the ability of bacteria and other microorganisms to respond to and survive external stresses. VIMSS was established in 2002 with funding from the U.S. Department of Energy Genomics: GTL Program for Rapid Deduction of Stress Response Pathways in Metal/Radionuclide Reducing Bacteria. LBNL is operated by the University of California for the U.S. Department of Energy.

The Virtual Institute for Microbial Stress and Survival (VIMSS, <http://vimss.lbl.gov>) funded by the Dept. of Energy's Genomics:GTL Program, is dedicated to using integrated environmental, functional genomic, and comparative sequence and phylogeny data to understand mechanisms by which microbes survive in uncertain environments while carrying out processes of interest for bioremediation and energy generation. To support this work, VIMSS has developed a Web portal with an underlying database and analyses for comparative functional genomics of bacteria and archaea. Since 2003 MicrobesOnline (<http://www.microbesonline.org>) has been enabling comparative genome analysis and currently includes 451 complete genomes and offers a suite of analysis and tools including: a multi-species genome browser, operon and regulon prediction methods and results, a combined gene and species phylogeny browser, a gene ontology browser, a workbench for sequence analysis (including sequence motif detection, motif searches, sequence alignment and phylogeny reconstruction), and capabilities for community annotation of genomes.

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

gene expression changes across multiple conditions and to search for similarities to these patterns. The information integration and analysis performed by VIMSS serves not only to generate insights into the stress responses and their regulation in these microorganisms, but also to document VIMSS experiments, allow contextual access to experimental data, and facilitate the planning of future experiments. VIMSS also is incorporating into MicrobesOnline publicly available functional genomics data from published research, so as to centralize and synergize data on microbial physiology and ecology in a unified comparative functional genomic framework.

117 ^{GTL}

Protein Complex Analysis Project (PCAP): Data Management and Bioinformatics Subproject

Adam P. Arkin,^{1,2,3} Ralph Santos,^{1,3} Y. Wayne Huang,^{1,3} Janet Jacobsen,^{1,2,3} Keith Keller,^{2,3} Steven S. Andrews,^{1,3} Steven E. Brenner,^{1,2,3} Max Shatsky,^{1,3} and **John-Marc Chandonia**^{1,3*} (JMChandonia@lbl.gov)

¹Lawrence Berkeley National Laboratory, Berkeley, California; ²University of California, Berkeley, California; and ³Virtual Institute for Microbial Stress and Survival, Berkeley, California <http://vimss.lbl.gov>

Project Goals: The Data Management and Bioinformatics component of the Protein Complex Analysis Project (PCAP) has two major goals: **1. to develop an information management infrastructure that is integrated with databases used by other projects within the Virtual Institute for Microbial Stress and Survival (VIMSS), and 2. to analyze data produced by the other PCAP subprojects together with other information from VIMSS to model stress responses relevant to the use of *D. vulgaris* and similar bacteria for bioremediation of metal and radionuclide contaminated sites. In addition to storing experimental data produced by the PCAP project, we will assess the quality and consistency of the data, and compare our results to other public databases of protein complexes, pathways, and regulatory networks. We will prioritize proteins for tagging, TAP, and study by EM based on analysis of VIMSS data and other bioinformatic predictions. All data we obtain on protein interactions will be analyzed in the context of the data currently stored in VIMSS. One of the primary goals of VIMSS is the creation of models of the stress and metal reduction pathways of environmental microbes. Ultimately, we wish to analyze PCAP data in such a way as to automatically generate hypothetical models of cellular pathways, which will be validated by comparison to experimental observations.**

The Data Management and Bioinformatics component of the Protein Complex Analysis Project (PCAP) has two major goals: **1. to develop an information management infrastructure that is integrated with databases used by other projects within the Virtual Institute for Microbial Stress and Survival (VIMSS), and 2. to analyze data produced by the other PCAP subprojects together with other information from VIMSS to model stress responses relevant to the use of *D. vulgaris* and similar bacteria for bioremediation of metal and radionuclide contaminated sites.**

We are developing a modular LIMS system to store data and metadata from the high-throughput experiments undertaken by the other PCAP subprojects. Each module of the LIMS corresponds to a step in the experimental pipeline. Modules for tracking bioinformatic data, tagged constructs, biomass production, tagless purification, and single particle EM data have been deployed. We are also developing WIST (Workflow Information Storage Toolkit), a set of libraries and tools for rapid

LIMS development. WIST allows LIMS programmers to design multi-step workflows using modular core components, which can be added and arranged through a simple, intuitive configuration and template mechanism.

We have also prioritized proteins for tagging, TAP, and study by electron microscopy based on analysis of gene expression data from the VIMSS Environmental Stress Pathway Project (ESPP) and bioinformatic predictions. To date, we have identified 403 *D. vulgaris* proteins as high-priority targets for tagging by the PCAP Microbiology Core. 127 of these proteins were chosen based on biological relevance (*e.g.*, involvement in sulfate reduction pathways) and analysis of ESPP data (*e.g.*, proteins for which the expression level is frequently observed to change in response to different stresses). In addition, *D. vulgaris* orthologs of *E. coli* proteins that were annotated as part of heteromeric or large (>250kD) homomeric complexes in the EcoCyc database or Butland TAP data were selected as targets. This was done in order to study the degree to which inter-protein interactions are conserved between orthologs, and to establish a baseline characterization of potential complexes to compare with the same proteins under stress conditions. 320 of the 403 genes are at or near the ends of their operons, and thus feasible to clone using our high throughput methods (further details are provided on Swapnil Chhabra's poster).

118 ^{GTL}

Gaggle: A Framework for Database Integration and Software Interoperability

Christopher Bare, Paul Shannon, Michael Johnson, and **Nitin S. Baliga*** (nbaliga@systemsbiology.org)

Institute for Systems Biology, Seattle, Washington

Project Goals: Molecular Assemblies Genes and Genomes Integrated efficiently: Characterize conserved protein complexes from a systems perspective.

MAGGIE Component 3

Collaborations for Gaggle implementation within MAGGIE

Co-PIs: Nitin S. Baliga; Mike Adams; Steven M Yannoni; Gary Siuzdak; John A. Tainer; Stephen R Holbrook

Institute for Systems Biology, Lawrence Berkeley National Laboratory (LBNL), The Scripps Research Institute, The Burnham Institute, University of Georgia (UGA), and University of California Berkeley (UCB)

A crucial challenge in systems biology is to combine the capabilities of diverse software tools and data resources to create an environment that promotes data exploration and analysis by a wide spectrum of users. A solution to this problem should recognize that data types, formats and software in this high throughput age of biology are constantly changing.

Gaggle is a simple, open-source Java software environment that helps to solve this problem of software and database integration. Guided by the classic software engineering strategy of *separation of concerns* and a policy of *semantic flexibility*, it integrates any new and *existing* popular programs and

web resources into a user-friendly, easily-extended environment by enabling sharing of four simple data types (names, matrices, networks, and associative arrays). Gaggle uses Java RMI and Java Web Start technologies and can be accessed at <http://gaggle.systemsbiology.net>. Gaggle is being routinely used as a mechanism for data sharing among the various components of the DOE-funded multi-institutional MAGGIE project.

119

Sensitivity Analysis on MS2 Viral Dynamics Using Interval Mathematics

Ozlem Yilmaz,* **Luke E. K. Achenie** (achenie@engr.uconn.edu), and Ranjan Srivastava

University of Connecticut, Storrs, Connecticut

Project Goals: Parameter determination is the most critical step in developing a model. The parametric errors introduced and the required precision for the parameters play an important role. The objective is to achieve parameter estimation using interval methods. This study represents the first step in accomplishing this goal.

Validated solutions of initial value problems (IVPs) can be obtained using interval analysis. A significant advantage over standard numerical methods is that an enclosure of the true solution is obtained. Since there are often measurement errors in an experiment, parameters determined based on the experimental data are also prone to errors. In this paper we employ a MATLAB version of the interval ODE solver VNODE (Nedialkov, 1999) to identify the most sensitive parameters in a biological model. The model under study explains how lytic RNA phage infects *Escherichia coli* C-3000 and the viral dynamics between the phage and its host at the intercellular level. Experimental data consisted of uninfected cell density (sensitive and resistant type), infected cell density, free phage density and substrate (glucose) concentration. There were 9 parameters determined experimentally and 6 parameters estimated using regression analysis (Jain et al., 2006). In our preliminary studies, each parameter was defined over an interval. Among all, the parameter corresponding to the rate of infection was found to be the most sensitive one. In the above studies, we note that the interval Hermite-Obreschkoff (IHO) method converges better than the interval Taylor series (ITS). The IHO unfortunately has high computational overhead and has convergence problems when the intervals are large. To alleviate these problems we have been investigating hybrid approaches that combine constraint satisfaction (Granvilliers et al., 2004 and Janssen et al., 2002) with IHO.

References

1. Nedialkov N., Computing rigorous bounds on the solution of an initial value problem, Ph.D. thesis, University of Toronto, 1999.
2. Jain R., Knorr A.L., Bernacki J. and R. Srivastava, Investigation of Bacteriophage MS2 Viral Dynamics Using Model Discrimination Analysis and the Implications for Phage Therapy, *Biot. Prog.*, 22:6, 1650-1658.
3. Granvilliers L., Cruz J. and P. Barahona, Parameter estimation using interval computations, *SIAM J. Sci. Comput.*, 26:2, 591-612.
4. Janssen M., Hentenryck P.V. and Y. Deville, A constraint satisfaction approach for enclosing solutions to parametric ordinary differential equations, *SIAM J. Numer. Anal.* 40:5, 1896-1939

120 ^{GTL}**The BioWarehouse System for Integration of Bioinformatics Databases**

Tom Lee, Valerie Wagner, Yannick Pouliot, and **Peter D. Karp*** (pkarp@ai.sri.com)

Bioinformatics Research Group, SRI International, Menlo Park, California

Project Goals: The goal of the BioWarehouse project is to allow scientists to integrate collections of DBs relevant to a genomics or systems-biology problem. BioWarehouse can integrate multiple public bioinformatics DBs into a common relational DB management system, facilitating a variety of DB integration tasks including comparative analysis, data mining, storage of locally generated data, and dissemination of data to the scientific community. All data are loaded into a common schema to permit querying within a unified representation.

BioWarehouse [1,2] is an open-source toolkit for constructing bioinformatics database (DB) warehouses. It allows different users to integrate collections of DBs relevant to the problem at hand. BioWarehouse can integrate multiple public bioinformatics DBs into a common relational DB management system, facilitating a variety of DB integration tasks including comparative analysis, data mining, storage of locally generated data, and dissemination of data to the scientific community. All data are loaded into a common schema to permit querying within a unified representation.

BioWarehouse currently supports the integration of the following databases: BioCyc, BioPAX protein interactions datasets (such as BIND), CMR, Eco2Dbase, ENZYME, Genbank (microbial subset), Gene Ontology, KEGG, MAGE-ML gene expression datasets, NCBI Taxonomy, and UniProt. Loader tools implemented in the C and Java languages parse and load the preceding DBs into Oracle or MySQL instances of BioWarehouse.

The BioWarehouse schema supports the following bioinformatics datatypes: chemical compounds, biochemical reactions, metabolic pathways, proteins, genes, nucleic acid sequences, features on protein and nucleic-acid sequences, gene expression data, protein interactions data, protein expression data, organism taxonomies, and controlled vocabularies.

BioWarehouse is in use by several bioinformatics projects. An SRI project is developing algorithms for predicting which genes within a sequenced genome code for missing enzymes within metabolic pathways predicted for that genome [3]. BioWarehouse fills several roles within that project: it is used to construct a complete and nonredundant dataset of sequenced enzymes by combining protein sequences from the UniProt and PIR DBs, and by removing from the resulting dataset those sequences that share a specified level of sequence similarity. Our current research involves extending the pathway hole filling algorithm with information from genome-context methods such as phylogenetic signatures, which are obtained from BioWarehouse thanks to the large all-against-all BLAST results stored within CMR. Another SRI project is comparing the data content of the EcoCyc and KEGG DBs using BioWarehouse to access the KEGG data in a computable form.

BioWarehouse is supported by the Department of Energy and by DARPA through the DARPA BioSPICE program for biological simulation.

References

1. BioWarehouse Home Page <http://bioinformatics.ai.sri.com/biowarehouse/>
2. T.J. Lee, Y. Pouliot, V. Wagner, P. Gupta, D.W.J. Stringer-Calvert, J.D. Tenenbaum, and P.D. Karp, "BioWarehouse: A Bioinformatics Database Warehouse Toolkit," *BMC Bioinformatics*, in press.

3. Green, M.L. and Karp, P.D., "A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases," *BMC Bioinformatics* 5(1):76 2004 <http://www.biomedcentral.com/1471-2105/5/76>.

121 ^{GTL}

A Cell Centered Database for Microbial Cells

Maryann E. Martone,^{1*} Joy Sargis,¹ Andrew McDonnell,² Gerry McDermott,² Carolyn Larabell,² Mark Le Gros,² Joshua Tran,¹ Willy Wong,¹ Vincent Ye,¹ **Harley McAdams**³ (hmcadams@stanford.edu), and Mark H. Ellisman¹

¹Center for Research in Biological Systems and the National Center for Microscopy and Imaging Research, University of California, San Diego, California; ²The National Center for X-ray Tomography, Lawrence Berkeley National Laboratory, Berkeley, California; ³Developmental Biology Department, Stanford University, Palo Alto, California

Project Goals: Development of high-throughput methods to identify and characterize spatially localized multiprotein complexes in bacterial cells.

The overarching objective of the Dynamic spatial organization of multi-protein complexes controlling microbial polar organization, chromosome replication, and cytokinesis project is to identify and characterize the regulatory proteins and protein/DNA complexes that control development of the bacterial cell and to determine the cellular locations where these molecules and complexes perform their function. This type of multi-disciplinary project employs a diverse array of methodologies and generates large amounts of diverse data the usefulness of which increases if easily shared between the collaborative sites. Moreover, forming a unified understanding of microbial cell biology requires integration of data from different experimental techniques into a common frame of reference. Without an informatics framework in which to deposit and search data, the process of building comprehensive structural maps of a bacterial cell type among collaborative groups becomes difficult. In a component of this Genomics:GTL project, we are establishing a distributed Cell Centered Database and associated informatics infrastructure to house the 3D tomographic data generated using electron and X-ray tomography, and correlated light microscopies. Databases have become integral parts of data management and data dissemination in biology. As the output of the microscopy community continues to increase, the utilization of databases for standard EM practice and large scale data sharing is becoming more critical. The Cell Centered Database (CCDB) was created to address the need for additional databases for cell level light and electron microscopic data, particularly geared towards large 3D datasets such as those produced by electron tomography. Adaptation of the CCDB to serve as a framework for microbial cell-centered projects is a natural extension of the CCDB's originally intended purpose.

The CCDB encompasses and hosts many types of data in the dimensional range that lies between gross morphology and macromolecular structure – the so-called “mesoscale”. Thus, unlike many structural databases on the web, the CCDB takes a “cell centered” approach in that it involves imaging of the cell by multiple techniques. These data and the software infrastructure that hosts and serves them are unique in scope, and provide the community with data that may have been technically difficult to obtain but is very rich in information content. This is particularly valuable resource for further examination, data mining and for use in development of computational models of structures and physiological processes that occur in cells and tissues. From its beginnings in 1998, the CCDB was envisioned as a grid and/or web-based federated database system. The CCDB pioneered

the production of a distributed, connectable repository system for managing and sharing data for a growing research community of microscopists.

During the past year, the CCDB completed a comprehensive set of data input forms, available through a secure grid-portal which can be launched from any web-browser. The portal architecture builds upon the Telescience framework (<http://telescience.ucsd.edu>), a grid-based portal architecture that simplifies and abstracts away the complexity of coordinating distributed resources, data, applications, and collaborations. Through the portal, users may upload, search and display their imaging data. Users manage their own groups and permissions, allowing them to share data with selected colleagues. Programmatic interfaces were created for microscopes so that data can be automatically acquired from the electron microscope and deposited into the CCDB. Similar interfaces are under construction for the X-ray microscope at LBNL and the laser-scanning multiphoton light microscopes at UCSD/NCMIR.

Although the original CCDB was created around imaging of eukaryotic cells, the CCDB models the process of reconstruction from a set of micrographs or images, including analysis and segmentation. Thus, although some of the tables were specific to eukaryotic organisms, most of the CCDB is generic for 2D, 3D and 4D microscopic imaging and was readily adapted for prokaryotes. Over the past year, we have been working with our collaborators on this Genomics:GTL project to modify tables such that they capture essential specimen preparation details for working with microbial cells. In addition, we have established a complete version of the CCDB at our partner site at Lawrence Berkeley National Laboratory. Tomography data generated from UCSD on *Caulobacter* are now housed in the CCDB. A newly constructed soft X-ray microscope at this site will be used to collect tomographic data of intact *Caulobacter* cells. The data model of the CCDB was designed to be extensible, to allow the incorporation of data from new microscopic imaging technologies such as soft x-ray tomography, by extending the CCDB schema.

The CCDB provides the backbone of a data management system for GTL microbial projects. Groups can either establish their own CCDB on site, or utilize the web-based portal version of CCDB hosted by UCSD. The ultimate goal is to allow the construction of representations of cells from data about their molecular constituents using a framework obtained from whole cell electron tomography, soft X-ray microscopy and high-resolution light microscopy. Future work will involve the development of spatial coordinate systems and ontologies to situate molecular constituents within their cellular contexts in order to bridge different scales of resolution and microscopy techniques.

122 ^{GT}L

Developments in the Systems Biology Workbench

Frank Bergmann,^{2*} Anastasia Deckard,² and **Herbert M. Sauro**¹ (hsauro@u.washington.edu)

¹Department of Bioengineering, University of Washington, Seattle, Washington and ²Keck Graduate Institute, Claremont, California

Project Goals: To develop a modular software framework (Systems Biology Workbench - SBW) for integrating the diverse software applications developed by the systems biology community.

In this abstract we describe some of the accomplishments achieved in 2006, both with respect to software and modeling. We have continued to improve the portability of the software across the three main platforms, Windows, Linux and Mac OS, to develop new modules and pursue integra-

tion with other third-party tools (Such as Copasi and Oscill8). In addition we have developed a new Wiki page where all documentation will reside (<http://sbw.kgi.edu/sbwWiki/>). To ensure that our software efforts are useful to the community we have also collaborated with three experimental groups to develop realistic biological models (Stem cell model, P53 model and MAPK model). The results of all three projects have been published.

Progress in Software Provision

We have made numerous minor improvements and some major changes in our software provision. Some of these developments are described in four papers we published this year. Here we only describe some of the major developments.

JDesigner: JDesigner (Visual editing tool) has undergone many improvements as a result of user feedback. The interface has been redesigned to include better visualization, better support for SBML, autolayout of networks and provision for storing multiple parameter sets within a given model.

Jarnac: Jarnac is our scripting tool for developing models. We have ported a 'lite' version of Jarnac to Linux and Mac OS, called JarnacLite, which permits users to describe a pathway very rapidly in text form, it can then be submitted to other SBW enabled editors such as JDesigner, Copasi or cellDesigner, where they can be refined.

3D Tool: The most exciting project we are working on is the 3D tool which allows network models to be visualized as 2D planes from which rise translucent pillars representing the concentrations of species. The visual representation can be fed in realtime from a SBW compatible simulator or data file so that the pillars will rise and fall as the model evolves. This gives a very interesting and different perspective on the model. In collaboration with Virginia Tech (VT), we are now able to visualize pathway simulations in real-time on the GigaPixel displays at VT. GigaPixel displays are physically large, high resolutions displays (>50 screens) that enable large quantities of data to be displayed simultaneously.

Composition Standard: We have published a proposal for a human readable format for building large models from small submodels. Current standards (SBML) can only describe a single model at a time. With the development of large models and in particular the rise of synthetic biology, there is a growing need to develop a compositional framework analogous to similar initiatives in the electronics field where standards such as Verilog have greatly stimulated the ability to share components and circuits (Details can be found at our Wiki: <http://sbw.kgi.edu/sbwWiki/>)

User Base: Our software continues to show increased usage, now running at around 600 downloads per month, with a total of 6500 downloads since the last GTL meeting. A recent review in Nature Biotechnology (2006) highlighted our GTL software as the best systems biology software currently available. On a number of occasions in 2006 we were ranked the number one bioinformatics project on source forge (out of ~700 projects).

123 ^{GTL}

The Ribosomal Database Project II: Introducing *myRDP* Space and Quality-Controlled Public Data

J. R. Cole* (colej@msu.edu), Q. Wang, B. Chai, E. Cardenas, R. J. Farris, A. M. Bandela, A. S. Kulam-Syed-Mohideen, D. M. McGarrell, G. M. Garrity, and **J. M. Tiedje**

Center for Microbial Ecology, Michigan State University, East Lansing, Michigan

Project Goals: The Ribosomal Database Project II (RDP) offers aligned and annotated rRNA sequence data and analysis service to the research community. These services help researchers with the discovery and characterization of microbes important to bioenergy production, biogeochemical cycles, and bioremediation.

Through its website (<http://rdp.cme.msu.edu>), The Ribosomal Database Project II (RDP) offers aligned and annotated rRNA sequence data and analysis service to the research community (Cole et al., 2006. Nucleic Acids Research; doi: 10.1093/nar/gkl889). These services help researchers with the discovery and characterization of microbes important to bioenergy production, biogeochemical cycles, and bioremediation.

Updated monthly, the RDP maintains 286,257 aligned and annotated quality-controlled rRNA sequences as of November 2006 (Release 9.45; Fig. 1). As a major quality improvement, all sequences are now tested for sequence anomalies, including chimeras, using Pintail from the Cardiff Bioinformatics Toolkit (Ashelford et al., 2005. Appl. Environ. Microbiol. 71:7724-7736).

***myRDP* Space:** This new feature allows users to maintain their own private sequence collection on the RDP servers aligned in sync with the RDP public alignment. With a *myRDP* account, researchers upload private rRNA sequences in Sequence Groups, which can range from a single sequence to thousands of sequences. Any combination of *myRDP* sequences and RDP public sequences can be selected for download or further analysis with the RDP tool suite. Sequences can be downloaded in formats ready for input to a wide variety of third-party phylogenetic and ecological tools.

After upload, *myRDP* sequences are automatically placed into the bacterial taxonomy using the RDP Classifier and aligned to match the RDP public alignment using RNACAD. Since the alignment remains in sync with the RDP public alignment, there is no need for the alignment compromises necessary to maintain compatibility between physically separated alignments. The RDP Classifier is accurate on sequences as small as 100 bases, although the information content varies along the molecule (Fig. 2).

***myRDP* Pipeline:** The new *myRDP* release incorporates a high-throughput sequence processing pipeline tailored to the requirements of single-read environmental sequence projects. The *myRDP* Pipeline consists of an integrated suite of publicly available and in-house developed programs. It provides the researcher with a simple path from sequencer output to quality-controlled, aligned sequences and analysis.

Video Tutorials: New short video tutorials demonstrate some of the more complex analytical tasks, including use of the new *myRDP* Pipeline. These tutorials average three minutes in length. They capture the screen as the tasks are performed, while the narrator explains the tasks and the choices available to the user.

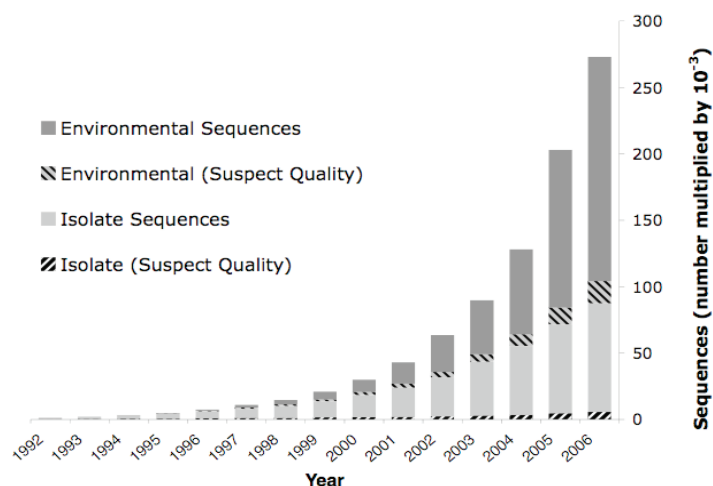


Figure 1. Increase in number of publicly available bacterial small-subunit rRNA sequences. Suspect quality sequences were flagged as anomalous by Pintail in testing with two or more reference sequences from different publications.

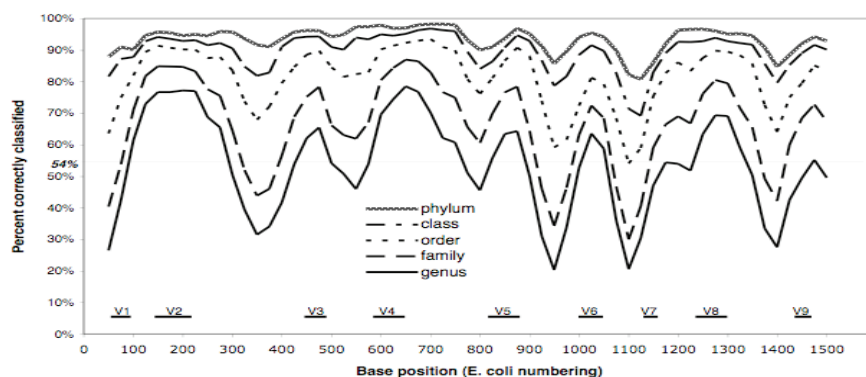


Figure 2. Classification accuracy rate for 16S rRNA sequence fragments of 100 bases. The gray bars on the x axis define the hypervariable regions. The V2 and V4 regions may make attractive targets for in-depth taxonomic analysis of environmental samples by the new short-read sequencing technologies.

124 ^{GTL}An Integrated Knowledge Resource for the *Shewanella* Federation

Nagiza F. Samatova^{1,*} (samatovan@ornl.gov), Denise Schmoyer,¹ Tatiana Karpinets,¹ Guruprasad Kora,¹ Sergey Passovets,¹ Michael Leuze,¹ and **Ed Uberbacher**¹ (ube@ornl.gov)

Collaborators from Shewanella Federation: Timothy S. Gardner², Gyorgy Babnigg³, Carol S. Giometti³, Margrethe Serres⁴, Anna Obratsova⁵, Grigoriy E. Pinchuk⁶, Alexander Beliaev⁶, Margaret F. Romine⁶, Kenneth Nealson⁵, and James K. Fredrickson⁶

¹Oak Ridge National Laboratory, Oak Ridge, Tennessee; ²Boston University, Boston, Massachusetts; ³Argonne National Laboratory, Argonne, Illinois; ⁴Marine Biology Laboratory, Woods Hole, Massachusetts; ⁵University of Southern California, Los Angeles, California; and ⁶Pacific Northwest National Laboratory, Richland, Washington

Project Goals: This project is a component of the *Shewanella* Federation and as such contributes to the overall goal of applying the tools of genomics, leveraging the availability of genome sequence for 18 additional strains of *Shewanella*, to better understand the ecophysiology and speciation of respiratory-versatile members of this important genus. To understand these systems the SF is using genome-based approaches to investigate *Shewanella* as a system of integrated networks; first describing key cellular subsystems — those involved in signal transduction, regulation, and metabolism — then building towards understanding the function of whole cells and, eventually, cells within populations. As a general approach, the SF is collectively employing complimentary “top-down” bioinformatics-based genome functional predictions, high-throughput expression analyses, and functional genomics approaches to uncover key genes as well as metabolic and regulatory networks. The “bottom-up” component employs more traditional approaches including genetics, physiology and biochemistry to test or verify predictions. This information will ultimately be linked to analyses of signal transduction and transcriptional regulatory systems and used to develop a linked model that will contribute to understanding the ecophysiology of *Shewanella* in redox stratified environments.

Critical to the success of the *Shewanella* Federation (SF) project and, arguably, to the overall success of the Genomics: GTL program is the sharing and integration of various types of information and data. The *Shewanella* Knowledge Base (<http://www.shewanella-knowledgebase.org/>) is a data and knowledge integration environment that allows *Shewanella* investigators (1) to capture, integrate and retrieve diverse ‘omics’ data for systems biology studies; (2) to navigate and superimpose information across gene-, protein-, expression- and pathway-levels; and (3) to perform com-

Experiment	Condition 1	Condition 2	Biological Replicate	Project	Contributor(s)
2D Gel	O2 limited	Aerobic	CR23	FedEx1	Carol Giometti (ANL), Gyorgy Babnigg (ANL)
Proteome MS	Anaerobic with fumarate	Aerobic	Averaged	FedEx1	Margaret Romine (PNL), Mary Lipton (PNL)
Proteome MS	Anaerobic with fumarate	Aerobic	CR23	FedEx1	Margaret Romine (PNL), Mary Lipton (PNL)
Proteome MS	Anaerobic with fumarate	Aerobic	CR24	FedEx1	Margaret Romine (PNL), Mary Lipton (PNL)
Proteome MS	Anaerobic with fumarate	Aerobic	CR25	FedEx1	Margaret Romine (PNL), Mary Lipton (PNL)
Proteome MS	O2 limited	Aerobic	Averaged	FedEx1	Margaret Romine (PNL), Mary Lipton (PNL)
Proteome MS	O2 limited	Aerobic	CR23	FedEx1	Margaret Romine (PNL), Mary Lipton (PNL)
Proteome MS	O2 limited	Aerobic	CR24	FedEx1	Margaret Romine (PNL), Mary Lipton (PNL)
Proteome MS	O2 limited	Aerobic	CR25	FedEx1	Margaret Romine (PNL), Mary Lipton (PNL)
Transcriptome Microarray	Aerobic with CaCl2	Aerobic	Averaged	FedEx1	Margaret Romine (PNL), Alexander Beliaev (PNL)
Transcriptome Microarray	Aerobic with CaCl2	Aerobic	CR23	FedEx1	Margaret Romine (PNL), Alexander Beliaev (PNL)
Transcriptome Microarray	Aerobic with CaCl2	Aerobic	CR24	FedEx1	Margaret Romine (PNL), Alexander Beliaev (PNL)
Transcriptome Microarray	Anaerobic with fumarate	Aerobic	Averaged	FedEx1	Margaret Romine (PNL), Alexander Beliaev (PNL)
Transcriptome Microarray	Anaerobic with fumarate	Aerobic	CR23	FedEx1	Margaret Romine (PNL), Alexander Beliaev (PNL)
Transcriptome Microarray	Anaerobic with fumarate	Aerobic	CR24	FedEx1	Margaret Romine (PNL), Alexander Beliaev (PNL)
Transcriptome Microarray	Anaerobic with fumarate	Aerobic	CR25	FedEx1	Margaret Romine (PNL), Alexander Beliaev (PNL)
Transcriptome Microarray	O2 limited	Aerobic	Averaged	FedEx1	Margaret Romine (PNL), Alexander Beliaev (PNL)
Transcriptome Microarray	O2 limited	Aerobic	CR23	FedEx1	Margaret Romine (PNL), Alexander Beliaev (PNL)
Transcriptome Microarray	O2 limited	Aerobic	CR24	FedEx1	Margaret Romine (PNL), Alexander Beliaev (PNL)
Transcriptome Microarray	O2 limited	Aerobic	CR25	FedEx1	Margaret Romine (PNL), Alexander Beliaev (PNL)
Transcriptome Microarray	O2 limited with fumarate	Aerobic	Averaged	FedEx1	Margaret Romine (PNL), Alexander Beliaev (PNL)

parative visual analyses in a cell systems context. The ultimate goal is to facilitate the generation of new hypotheses and knowledge about *Shewanella* systems behavior, while minimizing the researchers' effort, time and complexity.

The *Shewanella* Knowledge Base takes advantage of existing databases, resources and tools via direct linkages to avoid duplication of efforts occurring elsewhere. Its open architecture allows anyone interested to contribute and access information and data available for *Shewanella* species.

During the first year of the project the following progress has been made. We addressed the major problems associated with (a) data synthesis into a *common data model* from a distributed group of investigators, (b) data standards, and (c) experimental protocols representation. The system currently provides data models (compliant with community accepted data standards if available) for representing metabolic pathways, genome- and gene-level information, and experiment metadata. Initial schema design from the MIT ExperiBase project was extended to formally describe experimental campaigns, protocols, and computational analysis results for the following data types: mutants, cell culture, transcriptomics, and proteomics.

The project has also developed interfaces for datasets selection for comparative visual analysis of gene and protein expression profiles superimposed with publicly available biological information. For example, the investigator might be interested in identifying genes and subsystems (networks) required for anaerobic respiration of *S. oneidensis* MR-1 with various electron acceptors. The figures depict example questions the investigators may exploit, specifically:

- 1) What experimental data is available for FedEx1 campaign dealing with aerobic and anaerobic growth of *S. oneidensis* MR-1 under steady state conditions? (Fig. A)
- 2) How do proteomics and transcriptomics expression compare across gene clusters of interest? (Fig. B)
- 3) How do operon predictions by BioCyc match the expression predictions, or what proteins in the cluster are missed in proteomics experiments? (Fig. C)
- 4) How does the activity of enzymes involved in anaerobic and aerobic respiration change at the pathway level during the transition of *S. oneidensis* from aerobic to anaerobic growth? (Fig. D)
- 5) Are cytochrome mutants of interest available and what phenotypes do they have? (see web-site)

In the outgoing years the data model will be extended to support other types of SF-related data, various visual interfaces and analysis tools for comparative “omics” studies in the systems biology context. Capabilities for automatic data uploads and data access controls will be provided.

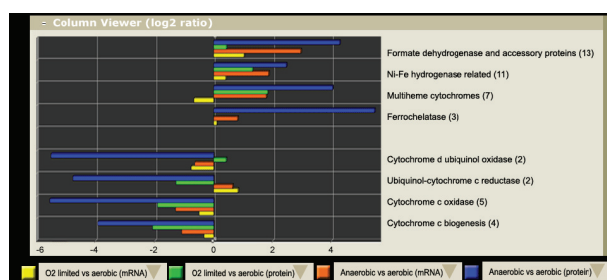


Figure B. Consistent changes in the gene and protein abundance of some respiration related enzymes under the same growth transition. Several mono- and di-heme cytochromes and enzymes involved in their biogenesis show down-regulation in anaerobic conditions versus aerobic ones. Several multi-heme cytochromes and related enzymes show consistent up-regulation.

Table Viewer (log2 ratio) : BioCyc Operons

Gene ID	Name					Function Description
SO2089	hypA	1.4	5.6	2.9	5.8	NiFe hydrogenase expression/formation protein, HypA
SO2090	hypE	2.2	3.8	2.1	4.1	NiFe hydrogenase accessory/formation protein, HypE
SO2091	hypD	0.9	2.7	2.5	3.4	NiFe hydrogenase expression/formation protein, HypD
SO2092	hypC	0.9	4.6	1.7	4.4	NiFe hydrogenase assembly chaperone, HypC
SO2093	hypB	-0.4	2.6	1.0	2.9	Ni-Fe hydrogenase assembly, Ni insertion (HypB)
SO2094	hypF	-1.3	-6.8	-0.6	-8.6	Ni-Fe hydrogenase assembly, Ni insertion (HypF)
SO2095	hyaE	0.0		1.7	4.7	NiFe hydrogenase assembly chaperone, HyaE
SO2096	hyaD	1.3		-1.2	4.7	NiFe hydrogenase maturase
SO2097	hyaC	0.9	1.6			quinone-reactive Ni/Fe hydrogenase, cytochrome b subunit (HyaC)
SO2098	hyaB	-1.3	3.4	0.3	4.2	quinone-reactive Ni/Fe hydrogenase, large subunit (HyaB)
SO2099	hyaA	-0.2	-0.9	2.9	-1.0	quinone-reactive Ni/Fe hydrogenase small subunit, HyaA

Legend:
 Yellow: O2 limited vs aerobic (mRNA)
 Green: O2 limited vs aerobic (protein)
 Orange: Anaerobic vs aerobic (mRNA)
 Blue: Anaerobic vs aerobic (protein)

Figure C. Consistency of expression for Ni-Fe hydrogenaze genes with the operon predictions by BioCyc. The predicted hypAED operon shows a consistent up-regulation both at gene and protein levels, while the other operon shows an overall up-regulation, but the hypF gene shows a significant down-regulation, also certain genes in the operon have missing protein expression values.

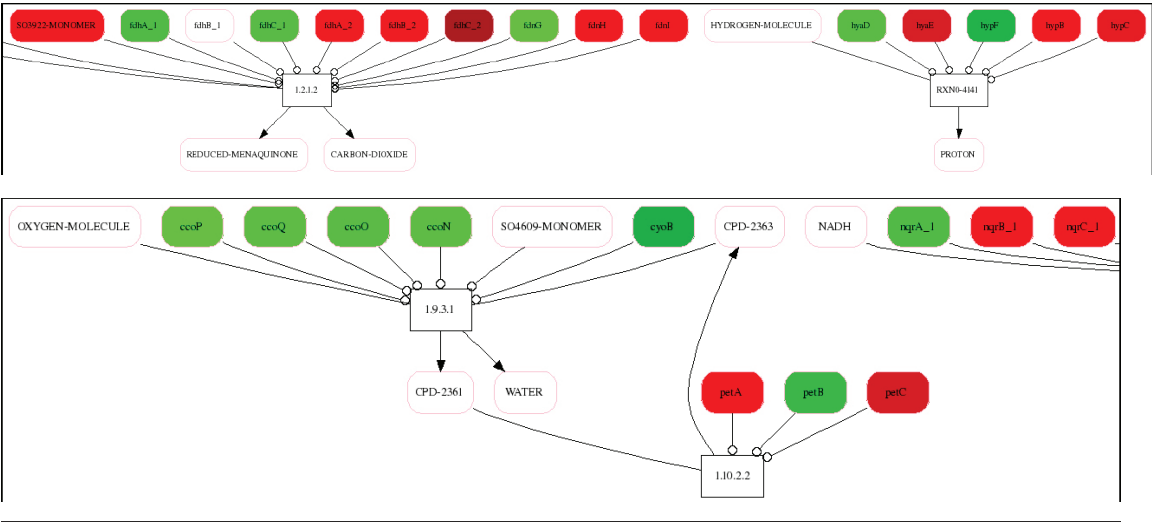


Figure D. The fragments of “anaerobic respiration – electron donors reaction list” (top) and “aerobic respiration – electron donor II” (bottom) pathways in ShewCyc from SRI with gene expression during the transition of *S. oneidensis* from aerobic to anaerobic growth.

This research sponsored by the U.S. DOE-BER, Genomics:GTL Program. Oak Ridge National Laboratory is managed by UT-Battelle, LLC, for the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

125 ^{GTL}**Informatics Strategies for Large-Scale Novel Cross-linker Analysis****Gordon A. Anderson**^{1*} (Gordon@pnl.gov), Nikola Tolic,¹ Xiaoting Tang,² and James E. Bruce²¹Environmental Molecular Science Laboratory, Pacific Northwest National Laboratory, Richland, Washington <http://www.emsl.pnl.gov> and ²Department of Chemistry, Washington State University, Pullman, Washington <http://www.wsu.edu/proteomics>

Project Goals: This project is focused on the development of novel cross-linker molecules to allow detection of protein-protein interactions. This work is done in collaboration between Washington State University (WSU) and the Pacific Northwest National Laboratory (PNNL). The researchers at WSU are developing the cross-linker molecules and methodologies while the PNNL researchers are focusing on the development of informatics tools to enable data analysis and identification.

The analysis of protein interactions in biological systems represents a significant challenge for today's technology. Chemical cross-linking provides the potential to impart new chemical bonds in a complex system that result in mass changes in the analysis of system tryptic peptides. However, system complexity and cross-linker product heterogeneity have precluded wide-spread chemical cross-linking use for large-scale protein interaction identification. The development of mass spectrometry identifiable cross-linkers called Protein Interaction Reporters (PIRs) has enabled on-cell chemical cross-linking experiments with product type differentiation. However, the complex datasets resultant from PIR experiments demands new informatics capabilities to allow interpretation. This presentation details our efforts to develop such capabilities and describes the program X-links which allows PIR product type differentiation. Furthermore, we also present the results from Monte Carlo simulation of PIR-type experiments to provide false positive identification rate estimates for the PIR product type identification through observed precursor and released peptide masses. Our simulations also provide false positive estimations based on accurate peptide mass measurements and database complexity. Overall, the calculations show a low rate of false positive identification of PIR product types due to random mass matching at under 8% at 10 ppm mass measurement accuracy. In addition this presentation illustrates the effect of database complexity on the PIR strategies ability to uniquely identify peptides using the constraints from this methodology. The PIR strategy includes a concept of development of a constrained protein database that increases the ability to uniquely identify cross-linked peptides and thus proteins. This presentation illustrates this methodology and quantifies the effect on unique peptide identification.

This research was supported by the Office of Science (BER), U.S. Department of Energy, Grant No. DE-FG02-04ER63924.

Communication

126 ^{GTL}

Communicating Genomics:GTL

Anne E. Adamson, Shirley H. Andrews, Jennifer L. Bownas, Sherry A. Estes, Sheryl A. Martin, Marissa D. Mills, Kim Nylander, Judy M. Wyrick, Anita J. Alton, and **Betty K. Mansfield***
(mansfieldbk@ornl.gov)

Oak Ridge National Laboratory, Oak Ridge, Tennessee

Project Web Site: genomicsgtl.energy.gov

Project Goals: Help build the critical multidisciplinary community needed to advance systems biology research for DOE energy and environmental missions and foster industrial biotechnology. The Genome Management Information System (GMIS) contributes to DOE Genomics: GTL program strategies and communicates key GTL scientific and technical concepts to the scientific community and the public. We welcome ideas for extending and improving communications and program integration to represent GTL science more comprehensively.

Accelerating GTL Science

For the past 18 years, we have focused on presenting information regarding DOE genome programs and the national and international Human Genome Project to a wide variety of audiences. Our goal always has been to help ensure that investigators could participate in and reap the genomic revolution's scientific bounty, new generations of students could be trained, and the public could make informed decisions regarding complicated genetics issues. Since 2000, GMIS has built on this experience to communicate about the Office of Science's Genomics:GTL program and its payoffs for DOE missions.

GTL systems biology research is a departure from traditional scientific methodology into a new territory of complexity and opportunity requiring contributions of interdisciplinary teams from the life, physical, and computing sciences and necessitating unprecedented integrative approaches. Because each contributing discipline has its own perspective, effective communication is highly critical to the overall coordination and success of GTL. Part of the challenge is to help groups speak the same language, from team and research-community building and strategy development through program implementation and reporting of results to technical and lay audiences. Our mission is to inform and foster participation by the greater scientific community and administrators, educators, students, and the general public.

Specifically, our goals center on accelerating GTL science and its applications. They include the following:

- Encourage information sharing, strategy development, and communication among scientists and across disciplines to accomplish synergies, innovation, and increased integration of knowledge. A new research community centered around the advanced concepts in GTL is emerging.

- Help reduce duplication of effort.
- Increase public awareness about the value of understanding microbial and plant capabilities important for solutions to national and global challenges.

Since 2004–2005, we have worked with DOE staff and teams of scientists to develop and disseminate the GTL roadmap. Tasks included helping to organize workshops, capture workshop output, and conduct the myriad activities associated with creating a technical document of the roadmap's size and importance. The science program described in this roadmap, a planning and program-management tool, was favorably reviewed by the National Academy of Sciences. In 2005 and 2006, as part of the bioenergy mission goal set forth in the GTL roadmap, we helped to organize the Biomass to Biofuels workshop and produce the report *Breaking the Biological Barriers to Cellulosic Ethanol*. The workshop and its output were jointly sponsored by the DOE Office of Science GTL program and Office of Energy Efficiency and Renewable Energy.

For outreach and to increase program input and grantee base, we identify venues for special GTL symposia and presentations by program managers and grantees. We also present the GTL program at meetings of such organizations as the American Association for the Advancement of Science, American Society for Microbiology, Society for Industrial Microbiology, American Society of Plant Biologists, American Chemical Society, National Science Teachers Association, National Association of Biology Teachers, and Biotechnology Industry Organization. Examples of other venues at which we present are the Plant and Animal Genomes meeting and the DOE Science Bowl.

We mail some 1600 packages of technical and educational material each month to requestors and furnish handouts in bulk to meeting organizers hosting genomics educational events. We continue to create and update handouts, including a primer that explores the impact of genomics on science and society and flyers on careers in genetics and other relevant issues. We supply educational materials in print and on the web site about ethical, legal, and social issues (called ELSI) surrounding the increased availability of genetic information.

All GTL publications are on the public web site, which includes an image gallery, research abstracts, and links to program funding announcements and individual researcher sites. Additional enhancements are being developed and implemented, including specific pages for DOE missions that will be impacted by the GTL program and its Bioenergy Research Centers. In addition to the GTL web site, we produce such related sites as Human Genome Project Information, Microbial Genome Program, Microbial Genomics Gateway, Gene Gateway, Chromosome Launchpad, and the CERN Library on Genetics. Collectively, our sites receive nearly 16 million hits per month. Over a million text-file hits are received each month during nearly 400,000 user sessions lasting 13 minutes—well above the average time for web visits. We are leveraging this activity to increase visibility for the GTL program.

This research is sponsored by the DOE Office of Science Office of Biological and Environmental Research.

Ethical, Legal, and Societal Issues

127 ^{GTL}

Science Literacy Project for Public Radio Journalists

Bari Scott* (bariscot@aol.com)

SoundVision Productions®, Berkeley, California

Project Goals: The Science Literacy Project is designed to give mid-career reporters and producers—both general assignment journalists and those who specialize in reporting on environmental, health, and technology issues—intense training via weeklong workshops basic science, science reporting and the creative use of radio to communicate science stories.

In this new era reporters have a greater responsibility than ever to explain science's accelerating advances and their profound social implications to the general public. While many journalists are still learning the basics of human genetics, reporters need to understand much more—from cell mechanics and the workings of DNA to the regulation of gene expression and the activity of proteins. Only then can they explain these crucial developments clearly and accurately to the public.

To help radio journalists meet this challenge, SoundVision Productions has established its Science Literacy workshops. These weeklong workshops are designed to give mid-career reporters and producers—both general assignment journalists and those who've specialized in reporting on environmental, health, and technology issues—intense training in both science reporting and the creative use of radio to communicate science stories. Given today's rapidly changing media environment, participants also learn the most effective ways to use new media from online reporting to podcasts.

Previous Science Literacy workshops were held in Boston and San Francisco, and a future workshop will be held in Austin, Texas. The goal of the workshops is to shrink the widening knowledge gap between the scientific community and the general public by increasing the quality and number of science stories on the radio. To achieve this, we hope to increase the number of reporters who can report accurately on complex scientific research and discoveries and their social implications.

SoundVision selects twelve public radio producers and reporters to attend each workshop. To compete for one of the slots, applicants must have contributed frequently to news or public affairs programs on national, regional or local public radio outlets and represent stations and national or regional programs that reach broad audiences. We place a high priority on including journalists from rural and minority-controlled stations and networks, and the recruitment and selection processes are designed to encourage ethnic, racial and gender diversity.

Each Science Literacy workshop is built around roughly twenty presentations by scientists, science journalists, scientific researchers and radio production professionals. Participants learn basic science, the ethics of science reporting, special techniques for presenting complex scientific content on radio, and the unique limitations and advantages of radio production. Workshops feature field trips and informal gatherings with scientists. The workshops explore the interactions of DNA, RNA, and proteins and the machinery of the cell; recent discoveries about the most basic elements of life; nonpathogenic microbes, and the interactions between genes and the environment. Workshops focus on ethics in the post-human-genome-project era, including new questions about the relationship

* Presenting author

between science and business, the impact of highly patented science on society and the risks and responsibilities of manipulating life. In the workshops, leaders teach participants how scientific and journalistic methods differ and show them how to interview scientists, explore new research, spot and handle scientific controversies and fact check stories on tight deadlines. As a result, participants return to their stations confident and excited about tackling complex scientific stories.

“I am a better reporter because of what I learned at that workshop,” one said. “I use what I learned there almost every day.”

We originally thought that the Science Literacy Project would conclude after a few workshops, but interest exploded beyond our greatest expectations. Word of our workshops has spread just as journalists have come to realize how much they need to learn about science reporting. As a result, SoundVision has received far more inquiries and applications than we can accommodate for the remaining workshop and we hope to expand the project.

At first, while reporters and producers clamored to take part, news directors and editors didn't seek the science literacy training. But in the face of accelerating scientific discoveries and controversies, more news directors and editors applied to the program and they made up one quarter of the participants at our last workshop. Having news directors and editors at the same workshop with producers and reporters has given each group a greater understanding of the other's needs.

The Science Literacy Project also includes a web site that provides transcripts and selected audio from the training sessions, “tip sheets,” and online resources. We continue to support participants in pursuing complex science stories for their communities by providing follow up teleconferences.

The project was evaluated by Rockman *et al*, a well-established San Francisco evaluation firm with expertise in evaluating media projects and assessing the impact of training on journalistic practice. Evaluations before the workshop helped us tailor presentations to participants' needs. During the workshops, in a new addition to our daily evaluations, we asked participants after each presentation to list the key scientific concepts they had just heard about to find out which scientific information stayed with them. In our review two months after the project we learned that news directors and editors wanted basic information about a broad range of subjects at the workshops, while producers and reporters preferred studying fewer issues in greater depth and we're working to meet both needs. We were also pleased to learn that many participants went back to their stations and gave their own workshops using some of our handouts. Most important, at the end of the Science Literacy workshops, participants returned to their home stations confident that they could handle complex scientific stories well.

128 ^{GTL}**The DNA Files®****Bari Scott*** (bariscot@aol.com)

SoundVision Productions®, Berkeley, California

Project Goals: The DNA Files will include five hour-long, nationally distributed public radio documentaries and five short features that will run on National Public Radio. The project will include outreach components that promote science journalism in public radio and the ethnic media, engage the public in series-related events, and contribute to science education and journalism across a broad spectrum of platforms.

SoundVision's highly acclaimed series *The DNA Files*® has demonstrated that complex scientific issues can be made clear and exciting to listeners with little science background. *The DNA Files*, hosted by John Hockenberry, will continue to show the importance of cutting-edge science in everyday life while expanding its audience to include more minority and rural listeners. We'll increase the series' impact in schools, museums, news outlets and beyond through our growing network of outreach services, media projects, and learning programs along with our expanding web site and proposed podcasts.

The DNA Files will include five hour-long, nationally distributed public radio documentaries and five short features that will run on National Public Radio. The series will explore revolutionary developments in toxicogenomics and individualized medicine, comparative genomics, neurogenetics, climate change, and food biotechnology. These complex and crucial subjects will be presented in a clear engaging style that makes them live for the general audience. Listeners will hear from scientists, government officials and corporate spokespeople as well as everyday citizens who have direct personal experience with the world of genetic research.

The DNA Files will also include outreach components that promote science journalism in public radio and the ethnic media, engage the public in series-related events, and contribute to science education and journalism across a broad spectrum of platforms. The outreach and education services include:

Media Support and Training

To help make science clear and relevant to a diverse population, *The DNA Files* will make a variety of resources available to journalists and media outlets. SoundVision will provide radio producers around the country with talk show discussion topics targeted to specific ethnic communities as well as general audiences and lists of experts whom reporters can use as sources for *DNA Files*-related programming and news reports. We'll also offer *The DNA Files Style Book* outlining best practice for scientific journalism online and send out short news alerts to reporters and editors to identify news stories related to *The DNA Files* 3 documentaries.

Finally, we will extend the series' impact by prompting local media outlets to create series-related stories and projects that are targeted to their specific audiences. To that end, SoundVision will work with stations to develop local programming that will position them as innovative science resources in their markets. At this point, twelve stations are taking part in the project, including Alabama Public Radio which plans to produce five sound-rich features for local broadcast during *Morning Edition*

and *All Things Considered* and five television news features produced for the local PBS station. They also plan to coordinate stories with local newspapers and develop a full-featured dedicated web presence, making sure the content and distribution methods of these stories are relevant to younger and minority audiences.

Educational Programs

San Francisco's world renowned Exploratorium Science Center will translate the documentaries' content into museum activities. The workshop activities will be disseminated to science centers around the country. The nationally respected museum of science education will also develop three interactive, hands-on teaching modules to extend the influence of *The DNA Files* beyond the air-waves. These modules will be available online and will also provide our station partners and other interested stations with a powerful, easy-to-use tool they can modify for their community programs.

Online Media: The Digital World

In response to the changing media landscape, *The DNA Files* is exploring new ways to expand our use of computer technology to create and distribute programs. To facilitate research, *The DNA Files* team has created an in house "Intranet" and uses Encode software to collect, organize and disseminate research materials for producers and staff. We are continuing to expand our information-packed multimedia web site by providing online "toolkits" to help reporters, editors, museum directors, teachers, and home-schooling parents build articles and lesson plans around *The DNA Files*. The site will include in-depth articles related to each of the five documentaries; background information and research for editors and reporters; a library of links to related web sites, and *The DNA Files Style Book*. *The DNA Files'* improved website will support public radio programming and museum and school programs that can stimulate public interest in science long after the series airs.

Evaluation

An independent firm will evaluate *The DNA Files* by conducting online user surveys and interviewing listeners to gauge their understanding and retention of the project's key themes. An evaluation of the original *The DNA Files* series concluded: "the style and format were highly effective in raising comprehension and awareness of the content among the focus group participants." *DNA Files* producers had, they said, "established an effective, appealing model for blending traditional and nontraditional public radio science formats with valuable awareness-building content."

The DNA Files has won numerous awards, including the George Foster Peabody Award, the Alfred I. DuPont-Columbia University Award, the American Association for the Advancement of Science Journalism Award, the Robert Wood Johnson Foundation Award, the American Institute of Biological Sciences Broadcast Award, and the Society of Professional Journalists Excellence in Journalism Public Service Broadcast Award.

USDA and DOE Joint Research

USDA and DOE Joint Research: Plant Feedstock Genomics for Bioenergy

129 ^{USDA-DOE}

Manipulation of Lignin Biosynthesis to Maximize Ethanol Production from *Populus* Feedstocks

Clint Chapple (chapple@purdue.edu), Richard Meilan, and Michael Ladisch

Purdue University, West Lafayette, Indiana

NON-TECHNICAL SUMMARY: High gasoline prices, global warming, national security, and the limitations of global petroleum resources have reinvigorated worldwide interest in renewable resources as a feedstock for liquid transportation fuels, particularly those derived from cellulose. As a perennial woody plant, hybrid poplar (genus *Populus*) offers several advantages with regard to cellulosic biofuel production including rapid growth rates, the ability to cycle nutrients, a wide geographic distribution, genetic diversity, amenability to genetic engineering, and abundant genomic resources. The phenolic cell wall polymer lignin constitutes a significant barrier to biomass conversion but, at the same time, it is essential to normal plant growth and development. Recent advances in our understanding of how lignin monomers are synthesized provide us with an opportunity to modify the content and composition of the lignin polymer. The research to be conducted will enable us to rationally assess the cost savings that could result from using genetically engineered poplar, instead of corn, as a feedstock for producing biofuels.

OBJECTIVES: 1) Generation of transgenic poplar up- or down-regulated for four enzymes known to impact lignin quantity and quality; 2) Development of metabolic profiling methods for poplar and their application to greenhouse- and field-grown wild-type and transgenic plants; 3) Morphometric analysis of transgenic lines grown in field plots; and 4) Cell wall deconstruction analysis of wild-type and lignin-modified transgenic lines.

APPROACH: Obj. 1) The expression of four enzymes in the lignin biosynthetic pathway will be up- and/or down-regulated. For each DNA construct, poplar cDNA will be synthesized from young shoot RNA using reverse transcriptase and PCR-amplified with gene-specific primers developed based on conserved regions within the genes' sequences identified from the poplar genome, the Arabidopsis genome, and other plant sequences. All constructs will be transformed into clone NM-6 (*Populus nigra* x *P. maximowiczii*) using an *Agrobacterium*-mediated transformation protocol. Obj. 2) Transformants will be tested for changes in lignin composition by a battery of lignin analyses (i.e., Klason lignin, pyrolysis GC-MS, and DFRC analysis). At the same time, HPLC and GC-MS will be used to assay total cell extracts and cell-wall hydrolysates from these plants to determine whether perturbations in phenylpropanoid pathway gene expression have led to alterations in free and/or cell wall-esterified phenolic compounds. Obj. 3) Morphometric analyses will be conducted on all lines in the field trial to ensure the transgenes have no deleterious effects on phenotype. All plants will be visually examined at least twice during the first and second growing seasons, including at least

once during leaf senescence in the fall and bud flush in the spring. Pairs of ramets with any unusual phenotypes will be photographed and measured for height, diameter, branch length, crown width, and changes in phenology. Obj. 4) Pretreatment will be carried out in batch-tube reactors. After pretreatment, an aliquot of the slurry in the tube will be collected and processed for image analysis, and the rest of the slurry in the tube will be processed by enzymatic hydrolysis using the NREL LAP009 enzyme digestibility procedure with minor modifications. The liquid separated from the solids in each condition will be filtered and analyzed via HPLC. These analyses will permit us to determine the impact of lignin modification on cell wall degradability.

130 ^{USDA-DOE}

Systematic Modification of Monolignol Pathway Gene Expression for Improved Lignocellulose Utilization

Richard Dixon (radixon@noble.org) and Fang Chen

Samuel Roberts Noble Foundation, Ardmore, Oklahoma

NON-TECHNICAL SUMMARY: Plant cell walls are made of three types of sugar polymer, cellulose, hemicellulose and pectin, and, as the cell wall develops, these are reinforced by lignin, a polymer of phenylpropane units (monolignols) that is recalcitrant to degradation. There are two stages involved in bioethanol production from lignocellulosic biomass: hydrolysis of the cell wall polysaccharides to their component hexose and pentose sugars, derived from cellulose/hemicellulose and hemicellulose respectively, and subsequent fermentation of the sugars to ethanol. The presence of lignin reduces access of enzymes and chemicals to hemicellulose and cellulose, thus reducing the efficiency of hydrolysis.

OBJECTIVES: The objectives of this proposal are 1) to determine which features of the lignocellulosic material (lignin content, lignin composition or other factors) are most detrimental to the fermentation of biomass to ethanol and 2) to develop the crop plant alfalfa (*Medicago sativa*) as a model system for genomic studies on biomass utilization..

APPROACH: Obj. 1. We have already generated transgenic alfalfa lines independently down-regulated in most (ten) of the enzymatic steps believed to be required for monolignol biosynthesis. Lignin content and composition have been determined in most of these lines (and cover a broader range of values than could be found in natural populations). The chemical analyses of the lignins will be completed, and the plant materials subjected to cell wall hydrolysis (acid and enzymatic) and fermentation. Yields of released sugars and bioethanol will be measured. We can then determine which features of the lignin polymer (content, composition, linkage types, etc) are most detrimental to sugar release and fermentation during bioethanol production, and design the optimal strategy for genetic modification of the plant feedstock for biofuel processing.

Obj. 2. Using genomic approaches (DNA microarray and informatics), we will discover additional genes necessary for lignin accumulation in alfalfa. These will be evaluated by down-regulation in transgenic plants as described above. We will develop approaches for non-biased discovery of genes impacting lignocellulose processing in *Medicago truncatula*, a model legume closely related to alfalfa, utilizing a large population of plants generated at the Noble Foundation that harbor transposon insertions. These lines will be screened for altered lignin properties by near infrared reflectance spectroscopy and simple staining procedures

Sorghum Biomass/Feedstock Genomics Research for Bioenergy**William Rooney**¹ (wlr@tamu.edu), John Mullet,¹ Steve Kresovich,² Doreen Ware,³ and P. Klein¹¹Texas A&M University, College Station, Texas; ²Cornell University, Ithaca, New York; and ³Cold Spring Harbor Laboratory, Cold Spring Harbor, New York

NON-TECHNICAL SUMMARY: Future production of renewable transportation fuels will require a consistent supply of biomass produced specifically for biofuel production. There will likely be many sources of biomass and species will be selected for their ecological fit, and their production and processing capability. Sorghum (*Sorghum bicolor* L. Moench) has the potential to be one of the species dedicated to biomass production because of its high productivity, drought tolerance, established production systems, and its genetic diversity. Research activities to develop sorghums for biomass production have been limited in the past. The purpose of this research is to use traditional and biotechnological approaches to produce sorghum genotypes with the genetic potential for use in bioenergy production.

OBJECTIVES: Within this grant period our specific objectives are to: (1) annotate genes, pathways and regulatory networks identified in the sorghum genome sequence that are important for biomass generation, and (2) identify, map and clarify the function of trait loci that modulate accumulation and quality of biomass in sorghum

APPROACH: (Obj. 1) Genes encoding proteins involved in biochemical pathways important for biomass generation, and plant composition related to biofuel production (i.e., starch, lignin, sugar, cellulose and hemicellulose), will be identified and projected onto biochemical pathways using the database MetaCyc. The pathway projections will provide a baseline of information on sorghum genes involved in biochemical pathways related to biomass/drought tolerance thus aiding our downstream analysis of QTL and traits. Moreover, the information on sorghum biochemical pathways in Gramene can be readily compared to information on other cereals and other organisms via Gramene's comparative mapping tools. This will help identify gaps in our current knowledge of sorghum biochemistry and help identify pathways and genes that may be useful to deploy in sorghum for biomass/bioenergy generation. (Obj. 2) The goals of objective 2 will be met in two approaches. First, grain, biomass, and carbohydrate yields will be measured in a population consisting of 175 recombinant inbred lines (RILs) ($F_{5,6}$) from the cross of BTx623 (a high yielding early flowering grain sorghum) $\times$ Rio (a high biomass sweet sorghum). Plant growth parameters will be analyzed to obtain a baseline for downstream meta-analysis. These include plant height, flowering time and tillering, traits that likely modulate carbohydrate partitioning in various tissues and total biomass. Traits that affect grain yield, biomass (i.e. the tissue harvest index and distribution of grain, stem, and leaf weight), the composition of structural and non-structural carbohydrates, and the overall energy gain of the plant will be evaluated. A genetic map of this population will be created and based on this map, QTL analysis will be carried out using QTL Cartographer, Mapmaker/QTL, or a similar analysis program. Second, tall, late flowering forage sorghum hybrids have the highest potential for total biomass generation. We are identifying genotypes that have high yield potential and excellent combining potential. In parallel, a RIL population will be developed by crossing tall, photoperiod sensitive late flowering genotypes that vary in biomass accumulation to explore the genetic basis of biomass accumulation and composition traits in this material. The population will be analyzed for variation in growth characteristics (growth rate and partitioning of growth) during the extended vegetative phase, for variation in lodging, total biomass and components of biomass

related to biofuel production as described above for the BTx623 × Rio population. This information will build a baseline of data on biomass production in tall, late flowering sorghums that have the highest potential for biomass production.

132 USDA-DOE

Streamlined Method for Biomass Whole-Cell-Wall Structural Profiling

John Ralph (jralph@wisc.edu), F. Lu, B. Sundberg, and S. Mansfield

U.S. Dairy Forage Research Center, USDA-Agricultural Research Service, University of Wisconsin, Madison, Wisconsin

NON-TECHNICAL SUMMARY: In wide-ranging research aimed at altering plant cell wall characteristics, by conventional breeding or modern genetic methods, one of the biggest problems is in delineating the effects on the cell wall. Plant cell walls are a complex conglomerate of a variety of polysaccharides and lignin. Each component alone is complex, and their interactions are only poorly characterized and understood. The most common approach has been to isolate and purify components and to characterize them in detail using a variety of methods. Such studies will always be necessary. As studies in which lignin-biosynthetic-pathway enzymes were targeted have made abundantly clear, simple compositional analysis is not sufficient. Some plants with only minor compositional changes have drastically altered chemical structure that belies the important alterations that can be made in processes ranging from natural digestibility in ruminant animals to industrial chemical pulping.

How can the structural components of the cell wall be readily characterized? Although other methods have their place, and can be more rapid (e.g. NIR), the difficulty in interpretation of some spectral methods, or the destruction of structure by chemical methods, assures that key features of cell walls benefiting, for example, biomass production and conversion are lost. A promising recent approach is the dissolution of the whole cell wall and NMR analysis. We intend to extend the methodologies to provide rapid structural profiling of plant materials, aiming for a “screening rate” of 20-30 samples per day. Such methodology will be useful to plant researchers worldwide.

OBJECTIVES: To provide the plant cell wall and biomass research communities with improved methods for polysaccharide and lignin structural profiling, based on complete cell wall solubilization and NMR. The aim is to develop and streamline procedures to allow 20-30 samples per day to be profiled.

APPROACH: The following improvements to the Dissolution/NMR method will be sought: a. Provide the necessary database, via model compounds and isolated components, to characterize component polysaccharides and lignins in whole-cell-wall mixtures. b. Optimize milling conditions for the various biomass sample types and seek alternative solutions that require less rigorous milling. c. Attempt to develop improved rapid dissolution methods that can be performed directly in the NMR tube; determine solvent systems that do not interfere with the correlation contours from polysaccharide and lignin components. d. Develop NMR methods that allow the crucial HSQC NMR spectra to be acquired in under 1 hour (on the whole cell wall sample). e. Develop methods for databasing and quantifying the 2D NMR cell wall spectra. f. With collaborators, attempt to develop chemometrics methods that can be applied to 2D NMR data.

KEYWORDS: lignin; plant chemistry; plant structure, polysaccharide, digestibility, chemical pulping, plant cell wall, NMR.

133 USDA-DOE

Development of a Proteoglycan Chip for Plant Glycomics

Chris R. Somerville (crs@stanford.edu)

The Carnegie Institute of Washington, Washington, D.C.

NON-TECHNICAL SUMMARY: In order to develop plants for use as optimal feedstocks for biofuels production from cellulosic biomass, it will be necessary to understand how the polysaccharides that comprise the majority of plant biomass are made and deposited in cell walls. More than a thousand genes for proteins that may be involved in synthesis and assembly of plant cell walls are evident in the genomic DNA sequences of the higher plants for which whole genome sequences are available. In order to assign functions to such a large number of genes, high-throughput assay methods will be required. This project concerns the development of a novel assay method that may facilitate the assignment of function to most of the relevant proteins. The goal is to develop reagents and methods that will allow presentation of a large number of different oligosaccharide structures on a solid surface in a small area. In principle, this can be accomplished by using robots to print small amounts of material at high density on a suitably surface-modified slide (eg., a “glycochip”) in much the same way that DNA microarrays are made. The glycans presented in this way are expected to serve as acceptors for glycosyltransferases, methylases and acetyltransferases, and as substrates for enzymes such as glycosidases that remove sugars from glycans. In this pilot project, we will focus on only those oligosaccharides that can be derived from plant proteoglycans containing O-linked glycans. The oligosaccharides will be prepared by using pure polysaccharide hydrolytic enzymes to fragment naturally occurring O-linked glycans which will then be purified by chromatographic methods. The glycochips produced in this way will be tested for their ability to act as acceptors in enzyme assays for glycosyltransferase enzymes from plants.

OBJECTIVES: To develop high throughput methods and reagents that will facilitate the assignment of function to large numbers of glycosyltransferases and other glycan modifying enzymes.

APPROACH: (1) Development of a series of transgenic plants that express synthetic peptides that become O-glycosylated in various ways and which have regions of sequence that facilitate purification of the glycopeptides; (2) purification of glycopeptides from transgenic plants; (3) determination of the structure of the glycans; (4) sequential cleavage of the glycans on each of the glycopeptides to produce a series of partial glycans; (5) production of glycochips by robotic spotting of the various glycopeptides onto chemically modified surfaces; (6) development of mass spectrometric methods for measuring the mass of glycopeptides in a microformat; (7) use of the glycochips to assay for glycosyltransferases activities in protein extracts from plants.

KEYWORDS: Arabinogalactan, AGP, Extensin, Hydroxyproline-rich glycoprotein, Glycosyltransferase, Glycomodule, O-linked glycan, cell wall,

134 ^{USDA-DOE}**Biochemical Genomics of Wood Formation: O-Acyltransferases for Alteration of Lignocellulosic Property and Enhancement of Carbon Deposition in Poplar****Chang-Jun Liu** (cliu@bnl.gov)

Brookhaven National Laboratory, Upton, New York

NON-TECHNICAL SUMMARY: The goal of GTL is to use newly acquired genomic data to better understand fundamental biological processes and enhance the translation of that scientific knowledge into new technologies for energy and environmental applications. Our project is going to characterize the *O*-acylation reactions participated in lignocellulosic biosyntheses using poplar genomics resources. *O*-acylation is involved in the cell-wall component polysaccharide and lignin biosyntheses. It affects the cell-wall sugar's solubility and the lignocellulosic bio-digestibility. *O*-acylation is also responsible for the structural modification and sequestration of a variety of polyphenolic metabolites required for wood formation. Understanding the mechanism of *O*-acylation at molecular level implicates biotechnological applications in genetic modification of lignocellulosic structures to facilitate biomass to bioethanol conversion, and in improvement of feedstock biomass production.

OBJECTIVES: 1) genome-wide identify acyl-CoA dependent acyltransferase genes from poplar genomics database; 2) systemically explore the tissue specific and stress-responsible expression patterns of *O*-acyltransferase genes to identify the enzymes specifically involved in lignocellulosic biosynthesis; 3) systemically characterize the biochemical functions of acyltransferases responsible for polysaccharide acetylation, lignol biosynthesis and phenolic compound modification.

APPROACH: Obj. 1. tblastn algorithm will be applied to search poplar genomics resources (*P. trichocarpa* V1.0, <http://genome.jgi-psf.org>) by using the highly conserved sequence motifs (HXXXD and DFGWG) of acyl-CoA dependent acyltransferases. In order to distinguish the potential soluble and membrane bound proteins, the encoded polypeptides of the identified gene candidates (at least ~50 gene models) will be subjected to the computational topology and post-translational modification analyses by using PSORT and SignalP web services to predict the protein sorting signal, subcellular localization site and the location of signal peptide cleavage sites in amino acid sequences. Obj. 2. The transcriptional profiling of putative acyltransferases will be analyzed both by "*in silico*" northern, based upon the high resolution poplar EST/microarray databases, and by QRT-PCR against mRNAs from different types of tissues (leaf, shoot, root, stem etc.) and the tissue sections representing different stages of developing wood, including the early expansion, late expansion, secondary cell wall formation, and programmed cell death (sapwood/heartwood) from poplars under normal growing conditions, the drought, salt stresses, and insect damage or physical wounding. Obj. 3. The recombinant proteins of the gene candidates that are highly expressed in wood-forming tissues will be produced using either *E.coli*, yeast, or *Drosophila* Gateway expression systems. Subsequently the combinatorial *in vitro* assays will be conducted by using different acyl-CoA donors and potential substrates including sugars, oligosaccharides, hydrolyzed and pre-deacylated pectin and xyloglucan, lignols, and other phenolics. Products detection and identification will be performed by LC-UV-ESI-MSⁿ, HPAEC-MS and MALDI-TOF-MS analyses.

135

USDA-DOE**Genomic Knowledgebase for Facilitating the Use of Woody Biomass for Fuel Ethanol Production****Vincent L. Chiang** (vincent_chiang@ncsu.edu)

North Carolina State University, Raleigh, North Carolina

NON-TECHNICAL SUMMARY: Situation and Problem: (A) Wood in forest trees is a major, potential lignocellulosic material for ethanol. (B) Trees can potentially be modified with a genome-wide approach for traits to overcome virtually any major biomass conversion barrier to ethanol production. (C) Gene expression and regulation of plant traits suited to ethanol production is poorly understood. Purpose: (A) This project examines, at the genome level, gene expression and regulation of lignocellulosic formation in *Populus trichocarpa*, a targeted energy tree crop. (B) The purpose of this project is to establish a knowledgebase about the possible genes and transcription factor genes involved in lignocellulosic formation and those genes that may enable effective manipulation of lignocellulosic traits to facilitate ethanol production.

OBJECTIVES: We propose a 3-year project to accomplish the following four objectives. (1) Chemical, biochemical and morphological profiling of TW development in *Populus*. (2) Oligo-microarray profiling of transgenics and TW development in *Populus*. (3) In vitro functional analysis of putative *Populus* xylan synthase genes. (4) Gene functional analysis in transgenic *P. trichocarpa*.

APPROACH: (1) Chemical, biochemical and morphological profiling of TW development in *Populus*: Wildtype and transgenics will be propagated for array characterization. For the TW system, we will profile cell wall trait changes at several different stages along the development of TW in Nisqually-1. These include cellulose, xylan, and lignin contents, lignin S/G ratios, xylan synthase activity, key lignin pathway gene transcript levels and enzyme activities, vessel/fiber ratios and TW fiber formation. These profiles reflecting changes due to preferential processes for the particular cell wall traits will guide microarray analyses to identify the involving genes and transcription factor genes and their contributions to these processes. (2) Oligo-microarray profiling of transgenics and TW development in *Populus*: RNAs from developing xylem of wildtype and selected transgenic *P. tremuloides* lines will be characterized by the updated full *Populus* transcriptome oligo-microarrays. RNAs from the developing xylem of the TW development stages with known cell wall trait/property profiles, will be characterized by the full genome microarrays. These RNAs will also be analyzed by miRNA oligo-microarrays designed with probes for detecting mature miRNAs that are mostly related to xylem development. Three biological replicates will be used in all array experiments. Differentially expressed genes will be determined and their transcript variation profiles between distinct transgenic levels or various TW developmental states will be correlated with the cell wall trait profiles to identify the proposed genes and genes encoding transcription regulators. (3) In vitro functional analysis of putative *Populus* xylan synthase genes: The array-selected and qRT-PCR confirmed putative xylan synthase genes will be expressed in our established *Drosophila* S2 cell system and the gene products will be characterized for biochemical functions. (4) Gene functional analysis in transgenic *P. trichocarpa*: We will select three transcription factor genes that may coordinate lignocellulosic accumulation and two miRNA genes that may regulate vessel and fiber cell development for overexpression in transgenic Nisqually-1 to determine their functions.

136 ^{USDA-DOE}**Genetic Dissection of the Lignocellulosic Pathway of Wheat to Improve Biomass Quality of Grasses as a Feedstock for Biofuels****Bikram Gill** (bsgill@ksu.edu) and Wanlong Li

Kansas State University, Manhattan, Kansas

NON-TECHNICAL SUMMARY: As fossil fuel reserves dwindle, we are about to transition from nonrenewable energy to renewable bioenergy. Demand for ethanol is increasing steadily as an alternative fuel as well as an octane-boosting and pollution-reducing additive to gasoline. Keeping in step with the demand will require higher quantity and quality of biomass. In the Great Plains, cereal crops (wheat, corn and sorghum), as well as native grasses (switchgrass, bluestem) predominate and are important but as yet relatively untapped resources for bioenergy.

OBJECTIVES: 1) Investigate the expression of ~80 candidate genes for lignin biosynthesis, their enzymatic activities, and lignin content and composition in different organs at different stages of diploid wheat plant; 2) Silence these 80 genes individually by VIGS; 3) Identify knockout mutants of these genes using TILLING and 4) Characterize the silenced tissues and knockout mutants by metabolite profiling.

APPROACH: obj. 1. Plants will be grown at seedling stage (Feekes stages F1-2), leaf sheath/stem elongation stages F4-5 and heading stages F8-9. The leaf, sheath, stem and spike tissues will be assayed for the expression of the 80 genes by QRT-PCR to decide the developmental phase and tissue on which all experiments will be done. Obj. 2. Conserved sequences will be used to silence all 80 genes individually by VIGS and the silenced tissues will be verified by QRT-PCR and positively silenced samples will be subjected to metabolite profiling, enzymatic assays and histochemical staining to determine the consequences of the genetic block in the lignocellulosic pathway. Obj. 3. We will do TILLING to screen for mutations for genes with major effect on the lignin content based on VIGS results. The homozygous plants containing genetic lesions for lignocellulosic pathway genes will be phenotyped with regard to growth and development, lignin content and composition, and lignocellulose degradability. Obj. 4. We will determine the lignin content of wheat silenced tissues, knockout mutants and controls by the acetyl bromide method, lignin composition by thioacidolysis methods, soluble and wall-bound phenolic compounds by HPLC, and polysaccharide content and composition by the phenol-sulfuric acid method. We will also histochemically characterize these plant materials by Mauns and Wiesner staining.

137

USDA-DOE**Using Association Mapping to Identify Markers for Cell Wall Constituents and Biomass Yield in Alfalfa****Charles Brummer**¹ (brummer@uga.edu), Kenneth J. Moore², and Jeff J. Doyle³¹University of Georgia, Athens, Georgia; ²Iowa State University, Ames, Iowa; and ³Cornell University, Ithaca, New York

NON-TECHNICAL SUMMARY: Alfalfa (*Medicago sativa*) is a potential biofuel crop because it produces high yield, its leaves can be used as a high value, high protein coproduct, it fixes atmospheric nitrogen, and it has beneficial effects on the environment. Improving alfalfa as a biofuel crop will entail breeding for increased biomass yield and altered cell wall composition. While traditional phenotypic selection can be successful, the perennial nature of alfalfa requires that a selection cycle lasts for several years. Decreasing the cycle time would increase genetic gain for all traits. This could be achieved using marker assisted selection for the traits of interest, but marker identification research conducted previously has not focused on representative alfalfa breeding populations nor has it examined wild germplasm as a source of new alleles to improve agronomically important traits. Our experiment will address these issues by studying both wild germplasm not typically used in alfalfa breeding programs and also a cultivated breeding population currently under selection. We will evaluate biomass yield and cell wall composition in the field. Concurrently, we will evaluate the genotype of each plant using genetic markers selected throughout the genome. We will also develop markers based on DNA sequence variation in genes of possible involvement in cell wall synthesis. Ultimately, this project will improve the efficiency of selection for enhanced bioenergy characteristics in alfalfa, produce numerous new markers at important candidate genes, and identify potentially useful alleles in wild germplasm.

OBJECTIVES: Our objectives are to use genomics approaches to identify chromosomal regions, and ultimately genes, controlling the two most important bioenergy traits, biomass yield and composition, and to develop genetic markers that can be used directly in applied plant breeding programs to improve the bioenergy qualities of alfalfa. We will pursue two complementary objectives to attain our goals: 1. Identify loci, and specific alleles, that control the concentration of alfalfa stem cell wall constituents and that are associated with biomass production using whole genome and candidate gene association mapping across a diverse set of natural diploid alfalfa accessions, and 2. Extend the analysis and methods used in the first objective to a tetraploid alfalfa breeding population currently under selection. As a result of this project, we (a) will have identified novel alleles in wild alfalfa germplasm that may be useful to improve cultivated alfalfa; (b) will have developed and used SNP markers in genes known to be involved in the biosynthesis of cell wall composition; (c) will be able to select individuals within a breeding population on the basis of these markers, and (d) will identify new alleles from wild germplasm useful for improving cultivated alfalfa. This experiment will provide the first estimate of linkage disequilibrium (LD) in alfalfa, both in a broad cross-section of wild diploid germplasm and in a practically important cultivated breeding population, both on a genome-wide and on an individual gene basis. Additionally, we will have applied association mapping to this important crop legume for the first time.

APPROACH: We will use association mapping to identify genome regions and candidate genes that are associated with biomass production and cell wall composition in both diploid and tetraploid alfalfa populations. We are proposing to begin by screening a broad diversity of diploid germplasm (three individuals from each of 96 plant introductions) in order to identify new genetic variation for

these traits that could be useful in alfalfa improvement. We will begin by analyzing diploid genotypes because they likely harbor a reservoir of unexploited genetic diversity and are more tractable for association mapping experiments than tetraploid genotypes. Subsequently we will extend the results to tetraploids. The tetraploid population we will examine is a breeding population currently under clonal selection at four locations, with 200 individuals being evaluated. As a breeding population, markers associated with traits could be immediately used in a recurrent selection program leading to the development of improved cultivars. Phenotypic analysis will be conducted based on field grown plant material clonally replicated to enable assessment of individual genotypes. In addition to biomass production and plant height measurements, we will conduct a thorough analysis of the stem cell wall composition of all entries. All plants will be genotyped throughout the genome with simple sequence repeat (SSR) markers, some of which will be selected based on their association with quantitative trait loci (QTL) for biomass yield, stem cell wall cellulose, hemicellulose, and lignin concentration, or agronomic traits that we have identified in other experiments. Concurrently, we will sequence portions of up to 100 genes that are candidate loci involved with cell wall biosynthesis. The sequencing will lead to the identification of single nucleotide polymorphisms (SNP), which we will develop into markers for those specific genes. All plants (288 diploid and 200 tetraploid) will be genotyped with the SNP markers. Association mapping will be conducted using the recently described mixed-model method that will account for underlying population structure within our two groups of genotypes (diploid and tetraploid), which will be analyzed separately. We will test for associations based on both genome-wide SSR molecular markers, as well as on SNP markers for 20 candidate genes, which will be developed from sequence data on 96 diploid and 20 tetraploid individuals.

Appendix 1: Participants

This section not available

Appendix 2: Web Sites

Program Web Sites

- Genomics:GTL Web site: <http://genomicsgtl.energy.gov>
- This book: <http://genomicsgtl.energy.gov/pubs/2007abstracts/>
- DOE Microbial Genome Program: <http://microbialgenomics.energy.gov>

Project and Related Web Sites

- Acid Mine Drainage Community Proteome Study
 - Proteogenomics of Communities: http://compbio.ornl.gov/biofilm_amd/
 - Peptide-resolved analysis of recombination in an AMD community: http://compbio.ornl.gov/biofilm_amd_recombination/
- BioWarehouse: <http://bioinformatics.ai.sri.com/biowarehouse/>
- Chisel: <http://compbio.mcs.anl.gov/CHISEL/>
- DOE Joint Genome Institute: <http://jgi.doe.gov>
 - JGI Eukaryote Genomes: <http://genome.jgi-psf.org>
- Environmental Molecular Science Laboratory: <http://www.emsl.pnl.gov>
- GAGGLE: <http://gaggle.systemsbiology.net>
- Invitrogen: <http://www.invitrogen.com>
- MAGGIE: <http://masspec.scripps.edu/MAGGIE/>
- Many Microbes Microarray Database (M3D): <http://m3d.bu.edu>
- Microbial Protein-Protein Interaction Database (MiPPI): <http://www.ornl.gov/sci/GenomestoLife/>
- populusDB: <http://www.populus.db.umu.se>
- ProRata: <http://www.MSProRata.org>
- Ribosomal Database Project II (RDP): <http://rdp.cme.msu.edu>
- *Shewanella* Knowledge Base: <http://www.shewanella-knowledgebase.org>
- Systems Biology Workbench: <http://sbw.kgi.edu>
 - Systems Biology Workbench Wiki: <http://sbw.kgi.edu/sbwWiki/>

- The SEED: <http://www.theseed.org>
 - The SEED genomic platform: <http://theseed.uchicago.edu/FIG/>
- The Institute for Genomic Research (TIGR): <http://www.tigr.org>
- Telescience: <http://telescience.ucsd.edu>
- UPSCBASE: <https://www.upscbase.db.umu.se>
- Virtual Institute for Microbial Stress and Survival: <http://vimss.lbl.gov>
 - Experimental Information and Data Repository: <http://vimss.lbl.gov/EIDR/>
 - MicrobesOnline: <http://www.microbesonline.org>
 - VIMSS Biofiles database: <http://vimss.lbl.gov/perl/biofiles/>

Author Index

A

Abulencia, Carl 28, 34, 89
 Achenie, Luke E. K. 183
 Adams, Michael W. W. 49, 51, 85, 126
 Adamson, Anne E. 195
 Adkins, Joshua N. 117, 128
 Adkins, Ronald M. 167
 Aklujkar, Muktak 35
 Allen, E.A. 111
 Allen, Eric E. 20
 Alm, Eric J. 10, 89, 92, 95, 100, 177, 180
 Alton, Anita J. 195
 Amat, Fernando 75
 Andersen, Gary L. 20, 26, 31
 Anderson, David J. 128
 Anderson, Gary 28, 89
 Anderson, Gordon A. 61, 193
 Anderson, Kevin K. 66
 Andersson, Anders 20
 Andrews, Shirley H. 195
 Andrews, Steven S. 52, 181
 Arkin, Adam P. 10, 11, 25, 52, 89, 92, 93, 95, 96, 97, 100, 101, 103, 177, 179, 180, 181
 Arrington, Ty 170
 Auberry, Deanna L. 66, 67, 68, 69
 Auberry, Kenneth J. 67, 128
 Auer, Manfred 52, 56
 Avila-Sakar, Agustin 52, 56
 Ayriss, Joanne 39

B

Babnigg, Gyorgy 114, 190
 Badger, Jonathan 170
 Baidoo, Edward 89, 91
 Baker, Brett J. 20, 111
 Baliga, Nitin S. 51, 151, 182
 Ball, David 56
 Banda, Philip 31
 Bandela, A. M. 188
 Banfield, Jillian F. 20, 109, 111, 113
 Bang, Duhee 41
 Bao, Gang 80
 Bare, Christopher 182
 Barrett, C.L. 146

Barry, Kerrie 25
 Barua, Soumitra 162
 Bauman, M. 132
 Beliaev, Alexander S. 114, 143, 152, 190
 Belnap, Chris 20
 Belov, Mikhail E. 128
 Benke, Peter 89, 91
 Bennett, George N. 157
 Bergmann, Frank 186
 Bertozzi, Carolyn 40
 Best, Aaron 134
 Betenbaugh, M. 132
 Biggin, Mark D. 52, 53
 Bonneau, Richard 151
 Borglin, Sharon 28, 89
 Bowman, Grant 76, 83
 Bownas, Jennifer L. 195
 Bradbury, Andrew 37, 38, 39, 40
 Brenner, Steven E. 52, 181
 Brettin, Tom 14
 Briñas, Raymond P. 78
 Bristow, James 25
 Britos, Leticia 123
 Brockman, Fred J. 10, 22
 Broderick, Joan B. 130
 Brodie, Eoin L. 10, 26, 28, 31, 89, 95
 Bruce, David 11, 25
 Bruce, James E. 61, 193
 Brummer, Charles 209
 Buchanan, Michelle V. 66, 67, 68, 69
 Buffleben, George M. 101, 103
 Bui, Olivia 139
 Burrows, E. H. 135
 Butland, Gareth 52, 53, 60
 Butler, J.E. 15

C

Callister, Stephen J. 117, 128
 Cannon, William R. 66, 67, 68
 Cao, Haishi 63
 Capone, Doug 32
 Cardenas, E. 188
 Carley, J. 26
 Carroll, S.L. 26
 Caspi, Ron 137
 Catarino, Teresa 44

Chai, B. 188
 Chain, Patrick 14
 Chakraborty, Romy 28, 89, 95
 Chan, Clara 113
 Chandonia, John-Marc 52, 60, 181
 Chaplen, F. W. R. 135
 Chapple, Clint 201
 Chasteen, Leslie 37, 38, 39
 Chatterjee, Ranjini 133
 Chavan, Milind A. 35
 Chen, Baowei 63
 Chen, E. 71
 Chen, Fang 202
 Chen, Wenqiong 100
 Chhabra, Swapnil 52, 58, 60, 89, 91
 Chiang, Vincent L. 207
 Chirica, Gabriela S. 89, 91, 101, 103
 Chivian, Dylan 10, 11, 89, 177, 180
 Cho, B.K. 146
 Chong, Song C. 105
 Ciccarone, V. 132
 Cipriano, Michael J. 89, 179
 Clark, M.E. 89, 95
 Clauss, Therese R.W. 128
 Clum, Alicia 14
 Cole, J.R. 12, 188
 Collart, Frank 62
 Collier, J. 171
 Collins, J.J. 153
 Colvin, Michael 175
 Comolli, Luis R. 75, 83
 Conlan, Sean P. 164
 Copeland, Alex 14
 Coppi, Maddalena 139
 Covalla, Sean F. 120
 Crane III, Edward J. 142
 Criddle, Craig S. 26, 105
 Cui, Qiang 176
 Culley, Dave 162
 Culley, David E. 10
 Cvetkovic, Aleksandar 49

D

Dai, Minghua 37, 38
 Daly, Don S. 66

Daugherty, Sean 170
 Davis, John 130
 DeBoy, Robert 170
 Deckard, Anastasia 186
 Deerinck, Thomas J. 83
 Dehal, Paramvir S. 11, 89, 177, 180
 DeJongh, Matthew 134
 Denef, Vincent J. 20, 109, 111
 Deng, Ye 25, 105
 Detter, Chris 25
 DeWitt, David 130
 DiBartolo, G. 111
 DiBartolo, Genevieve 20
 DiDonato, Laurie N. 139, 167
 DiDonato Jr., Raymond 120, 139
 Ding, Chris 84
 Dixon, Richard 202
 Doktycz, Mitchel J. 67, 69
 Donahoe, Samuel M. 151
 Dong, Ming 52, 53
 Downing, Kenneth H. 52, 56, 75
 Doyle, Jeff J. 209
 Driscoll, Mike E. 116, 153
 Drury, Elliot C. 89, 91, 97
 Du, Xuixia 117, 128
 Dubchak, Inna 89, 179
 Dubini, Alexandra 146
 Duke, Norma 44
 Durek, Thomas 41
 Duschene, Kaitlin 130
 Dye, N. 73

E

Elias, Dwayne A. 52, 58, 60, 89, 97
 Elkins, James 122, 146
 Ellisman, Mark H. 75, 83, 185
 Ely, R. L. 135
 Eppley, John M. 20, 111
 Erickson, Brian 109
 Estes, Sherry A. 195

F

Facciotti, Marc T. 151
 Fahland, Tom 139
 Faith, J.J. 153, 154
 Famili, Iman 139, 143, 145
 Fan, Zhanmin 133
 Farris, R. J. 188
 Fero, M. 76
 Fields, Matthew W. 25, 26, 28, 89, 91, 95
 Finzi, Juliette 32
 Fisher, Hugh 37, 38

Fisher, Susan J. 52, 53
 Foerster, Hartmut 137
 Foote, Linda J. 67, 69
 Fredrickson, James K. 8, 61, 114, 143, 145, 162, 190
 Fredrickson, Jim K. 117, 128
 Fulcher, Carol 137
 Fuss, Jill 51

G

Gabster, J. 11, 89
 Gaietta, Guido M. 83
 Gao, Haichun 162
 Garcia, Joseph 34
 Garczarek, Florian 52, 56
 Gardner, Timothy S. 153, 154, 190
 Gardner, Tim S. 116
 Garrity, G.M. 12, 188
 Gassman, Natalie 160
 Gates, Zachary 41
 Gaucher, Sara P. 89, 91, 101, 103
 Gelfand, Mikhail S. 156, 179
 Geller, Jil T. 28, 52, 58, 89
 Gentry, Terry J. 25, 26, 105
 Geydebekht, O. 143
 Ghirardi, Maria L. 146
 Gihring, Thomas M. 10
 Giles, B. 89, 95
 Gill, Bikram 208
 Giometti, Carol S. 8, 114, 190
 Gitai, Z. 71, 73
 Giuliani, Sarah 62
 Glaeser, Robert M. 52, 56
 Glaven, Richard 122, 165
 Goltsman, Daniela 20, 111
 Goltsman, Eugene 14
 Gopalan, B. 22
 Gorfien, S. 132
 Green, Michelle 137
 Grossman, Arthur R. 16, 146
 Gu, Baohua 105

H

Hadi, Masood 89, 91
 Hainfeld, James 78
 Hall, Evan T. 142
 Hall, Steven C. 52, 53
 Hammel, Michal 85
 Han, Cliff 14
 Harrison, S.H. 12
 Harwood, C. S. 124
 Harzman, Christina 130
 Hauser, Lauren 34

Hauser, Loren 174
 Haveman, Shelley 122
 Hayete, B. 153
 Hazen, T. C. 92, 95
 Hazen, Terry C. 10, 11, 25, 26, 28, 52, 58, 89, 93, 97, 100, 101, 103, 105, 107
 He, Qiang 89, 91, 95, 100
 He, Zhili 89, 91, 93, 95, 96, 100, 105, 162
 Hemme, Christopher L. 25, 89, 91, 93, 162
 Hemming, Christi 130
 Hendrickson, E. 28, 89
 Herrgard, M.J. 146
 Herring, Chris 122
 Herzberg, Osnat 43
 Hettich, Robert L. 20, 109, 111, 113, 124
 Hillesland, Kristina L. 11, 89, 98
 Hillson, N. 76
 Hixson, Kim K. 117, 120, 128
 Ho, Sam On 160
 Hoeprich, Paul 31
 Holbrook, Stephen R. 51, 84
 Holman, Hoi-Ying 28, 52, 58, 89, 95
 Holmes, Dawn E. 35
 Hong, S. 76
 Hood, Leroy 151
 Hooker, Brian S. 66, 67, 69
 Horowitz, Mark 75, 171
 Hu, Minghui 78
 Huang, Katherine H. 89, 92, 97, 100, 177, 180
 Huang, Y. Wayne 52, 89, 177, 180, 181
 Huitema, E. 71
 Hura, Greg 49, 85
 Hurst, Gregory B. 66, 67, 68, 124
 Hutcheon, Ian D. 31, 32
 Hutchinson, Don 34
 Hwang, C. 26, 89

I

Izallalen, Mounir 139

J

Jacobsen, Janet 28, 52, 89, 96, 177, 180, 181
 Jaitly, Navdeep 128
 Jap, Bing K. 52, 53

Indexed by page number.

Jardine, Phil M. 26, 105
 Jeans, Christopher 109, 113
 Jenney Jr., Francis E. 49, 85
 Jensen, Grant 80
 Jin, Jian 52, 53
 Joachimiak, Marcin P. 89, 177, 180
 Johnson, Erik 41
 Johnson, Jessica P. 120
 Johnson, Michael 182
 Jones, J. 132
 Joyner, Dominique C. 11, 28, 52, 58, 89, 100, 101, 103
 Juárez, Katy 165
 Judd, D. 132
 Juhn, F.S. 153

K

Kaipa, Pallavi 137
 Kale, Pat 14
 Kalisiak, Ewa 49, 126
 Kalisiak, Jarek 126
 Kalnejais, Linda 20, 111
 Karp, Peter D. 137, 184
 Karpinets, Tatiana 190
 Karpowicz, Steven 16
 Kaur, Amardeep 151
 Kazakov, Alexei E. 179
 Ke, Haiping 165
 Keasling, Jay D. 52, 58, 60, 89, 91, 93, 103
 Kehoe, John 40
 Keller, Keith 52, 89, 177, 180, 181
 Keller, Martin 28, 34, 89, 100, 107
 Kennedy, D. 143
 Kennel, Stephen J. 67
 Kent, Stephen 41
 Kiebel, Gary R. 128
 Kim, Arnold 175
 Kim, Byoung-Chan 165
 Kim, Sang-Hoon 130
 Kim, Younggyu 160
 Kiss, Csaba 37, 38
 Klappenbach, Joel 145
 Klein, P. 203
 Klonowska, A. 95
 Knight, E.M. 146
 Koenig, Peter 176
 Kora, Guruprasad 124, 190
 Kostantinidis, Konstantinos T. 8
 Kozina, Carrie 101, 103
 Krag, S. 132
 Kresovich, Steve 203

Indexed by page number.

Krummenacker, Markus 137
 Krushkal, Julia 167
 Kulam-Syed-Mohideen, A. S. 188
 Kulp, D. 15
 Kuske, Cheryl 34
 Kusnetsova, Larissa 78

L

Ladisch, Michael 201
 Land, Miriam 8, 116, 174
 Landorf, Elizabeth 62
 Lankford, Trish K. 67
 Lapidus, Alla 10, 14
 Larabell, Carolyn A. 86, 185
 Larimer, Frank 174
 Latendresse, Mario 137
 Lawrence, Albert 75
 Lawrence, Charles E. 164
 Leang, Ching 165, 167
 LeDuc, Phil 80
 Lee, Tom 184
 Lee, Y.C. 132
 Lefsrud, Mark 109
 Leigh, J. 28, 89
 Leuze, Michael 190
 Le Gros, M.A. 86
 Le Gros, Mark 185
 Li, Huilin 78
 Li, Jun 158
 Li, Wanlong 208
 Liao, James C. 165
 Lie, T. 28, 89
 Liebich, Jost 105
 Lightstone, Felice 175
 Lilburn, T.G. 12
 Lin, Chiann-Tso 69
 Lin, Li-Hung 10
 Lipton, Mary S. 8, 116, 117, 120, 123, 128
 Liu, Chang-Jun 206
 Livesay, Eric A. 69, 128
 Lo, I. 109, 111
 Londer, Yuri Y. 44
 Lou, Jianlong 40
 Lovley, Derek R. 15, 35, 120, 122, 139, 141, 165, 167, 170
 Lowry, Steve 10, 14
 Lu, F. 204
 Lu, Tse-Yuan S. 67
 Luna-Chavez, Cesar 85
 Lyman, Elena S. 78

M

Maguire, B.M. 86
 Mahadevan, Radhakrishnan 139, 141, 167
 Malik, Jitendra 52, 56
 Maltsev, Natalia 4, 22, 116
 Mansfield, Betty K. 195
 Mansfield, S. 204
 Mao, Fenglou 149
 Marks, James D. 40
 Marsh, Terence L. 130
 Marshall, Matt 8
 Martin, Sheryl A. 68, 195
 Martone, Maryann E. 185
 Masol, A. 26
 Mayampurath, Anoop 128
 Mayer, M. Uljana 63
 McAdams, Harley H. 75, 76, 83, 123, 171, 185
 McCrow, John P. 7
 McCue, Lee Ann 8, 116, 164
 McDermott, Gerry 86, 185
 McDermott, Jason E. 66
 McDonald, W. Hayes 66, 67, 68, 69, 124
 McDonnell, Andrew J. 86, 185
 McGarrell, D. M. 188
 McKeown, Catherine K. 67
 Meilan, Richard 201
 Menon, Angeli Lal 49, 85, 126
 Merchant, Sabeeha 16
 Methé, Barbara 167, 170
 Mielke, M. 73
 Milewski, Paul 176
 Miller, Lisa M. 46
 Mills, Marissa D. 195
 Mitchell, Julie C. 176
 Monroe, Matthew E. 128
 Moore, Kenneth J. 209
 Moore, Ronald J. 128
 Morita, Hiro 126
 Morrell-Falvey, Jennifer L. 67, 69
 Moser, Duane P. 10
 Mottaz, Heather M. 128
 Moul, John 43
 Moussavi, Farshid 75
 Mukhopadhyay, Aindrila 28, 52, 58, 60, 89, 91, 93, 103
 Mukhopadyay, Aindrila 97
 Mullet, John 203
 Munoz, Denise 51
 Mus, Florence 146
 Muske, Gerhard R. 61

Mylenski, Michael 78

N

Nealson, Kenneth H. 3, 7, 8, 142, 162, 190
 Nevin, Kelly P. 120, 122, 170
 Newberg, Lee A. 164
 Nicora, Carrie D. 117, 128
 Nie, Shuming 80
 Nogales, Eva 52, 56
 Norbeck, Angela D. 117, 128
 Nordstrom, D. Kirk 20
 Núñez, Cinthia E. 167
 Nylander, Kim 195

O

O'Neil, Regina A. 35
 Obratzova, Anna 8
 Obratzova, Anna 3, 190
 Oda, Y. 124
 Onstott, Tullis C. 10
 Orton, Daniel J. 69, 128
 Osterman, Andrei 3, 116, 156
 Owens, Elizabeth T. 67, 69

P

Paley, Suzanne 137
 Palsdottir, Hildur 52, 56
 Palsson, Bernhard O. 122, 146
 Palzkill, Timothy 162
 Pan, Chongle 109, 124
 Pan, Min 151
 Paša-Tolić, Ljiljana 128
 Passovets, Sergey 190
 Pavlik, Peter 37, 38, 39, 40
 Peeples, Jeanette 167
 Pelletier, Dale A. 66, 67, 68, 69, 124
 Pentelute, Brad 41
 Peppler, Terese 62
 Perez, E.X. 26
 Pesavento, Emanuele 37, 38
 Pessanha, Miguel 44
 Petritis, Kostantinos 128
 Pett-Ridge, Jennifer 31, 32
 Phan, Richard 28, 89, 101
 Pharkya, Priti 139
 Piceno, Y.M. 26
 Pinchuk, Grigoriy E. 114, 143, 145, 190
 Pincus, Z. 73
 Pingitore, Francesco 89, 91
 Podar, Mircea 34

Pokkuluri, P. Raj 44
 Poole, Farris L. 126
 Poole, Ferris 85
 Poole II, Farris L. 49
 Posewitz, Matthew C. 146
 Postier, Brad 167
 Postier, Bradley 139
 Pouliot, Yannick 184
 Pratt, Lisa M. 10
 Price, Morgan N. 11, 89, 177, 180
 Prior, David C. 128
 Pritchard, S. 71
 Prochnik, Simon 16
 Puljic, Marko 167
 Purvine, Samuel O. 117, 123, 128

Q

Qian, Luping 78

R

Ralph, John 204
 Rambo, Robert 85
 Rasmussen, Jytte 40
 Ravcheev, Dmitry 179
 Raymond, Jason 113
 Redding, Alyssa M. 89, 91, 97, 103
 Reed, J. 143
 Reed, Jennifer L. 145
 Reed, Samantha B. 143, 162
 Reiss, David J. 151
 Remis, Jonathan P. 52, 56
 Rhee, Sue 137
 Richardson, Paul M. 10, 11, 14, 25
 Ringbauer Jr., Joseph A. 89, 96
 Risso, Carla 139
 Rodionov, Dmitry 3, 156
 Rodrigues, Jorge L.M. 8
 Rodriguez, A. 4, 22
 Rokhsar, Dan 16
 Romine, Margaret R. 3, 4, 8, 114, 116, 117, 143, 145, 156, 158, 162, 190
 Rooney, William 203
 Rubin, Eddy 25
 Ryan, Kevin 78

S

Sachidanandam, R. 154
 Saffarini, Daad A. 152
 Salgueiro, Carlos A. 44
 Samatova, Nagiza F. 190, 124
 Santos, Ralph 52, 181

Sapra, Rajat 89, 91, 101, 103
 Sargis, Joy 185
 Sauro, Herbert M. 186
 Saxman, P.R. 12
 Schadt, Christopher W. 89, 105, 107
 Schiffer, Marianne 44
 Schilling, Christophe H. 139, 145
 Schmoyer, Denise D. 66, 67, 68, 190
 Schmutz, Jeremy 14
 Scholten, Johannes C. 145
 Schrader, P. S. 135
 Schwarz, Fred 43
 Scott, Bari 197, 199
 Scott, Joseph 49, 143
 Seibert, Michael 146
 Serres, Margrethe H. 8, 116, 190
 Shah, Manesh B. 20, 67, 68, 109, 111
 Shanks, Jacqueline V. 148
 Shanmukh, Saratchandra 49
 Shannon, Paul 151, 182
 Shapiro, Lucy 71, 73, 76, 83, 123, 171
 Sharp, Julia L. 66
 Shatsky, Max 52, 56, 181
 Shen, X. 171
 Shen, Yufeng 128
 Shi, Liang 61, 63
 Shirodkar, Sheetal 152
 Shukla, Anil K. 128
 Shutkin, Amy 89
 Simmons, Sheri 20
 Simons, Julie 176
 Singer, Mary 52, 58
 Singer, Steven W. 109, 113
 Singh, Anup K. 89, 91, 101, 103
 Singhal, Mudita 66
 Siuzdak, Gary 49, 51, 126
 Smith, Harold 43
 Smith, Richard D. 117, 123, 128
 Smith, Thomas M. 164
 Somerville, Chris R. 205
 Southam, Gordon 10
 Squier, Thomas C. 63
 Srivastava, Alok 151
 Srivastava, Ranjan 183
 Stahl, David A. 11, 28, 89, 96, 98, 103, 104
 Stedman, Kenneth 51
 Stolyar, S.M. 104
 Stolyar, S.S. 96
 Stolyar, Sergey M. 89

Stolyar, Sergey S. 11, 28
 Studier, F. William 46
 Sulakhe, D. 4
 Summers, Zarath 122
 Sun, Hui 14
 Sun, Jun 139
 Sundberg, B. 204
 Sutherland, John C. 46
 Syed, M. 4, 22

T

Taghavi, Safiyh 17
 Tainer, John A. 47, 49, 51, 85
 Tang, Xiaoting 61, 193
 Tang, Yinjie 89, 91
 Taverner, Tom 123
 Taylor, Ronald C. 66
 Thelen, Michael P. 20, 109, 111, 113
 Theriot, J. 73
 Thieman, S.B. 95
 Thompson, William A. 164
 Tiedje, James M. 8, 25, 130, 162, 188
 Tissier, Chris 137
 Tolić, Nikola 61, 128, 193
 Tolmachev, Aleksey V. 128
 Tomiya, N. 132
 Tonnessen, C.A. 86
 Torbeev, Vladimir 41
 Toro, Esteban 123
 Torok, Tamas 52, 58
 Tran, Joshua 185
 Trauger, Sunia A. 49, 126
 Trong, Stephan 14
 Tsutakawa, Susan 85
 Turse, Joshua 117
 Typke, Dieter 52, 56
 Tyson, Gene 20, 111

U

Urbacher, Ed 190
 Udelhoven, Rachel 130
 Udseth, Harold R. 128
 Ueki, Toshiyuki 165, 167
 Ulrich, Luke E. 155

V

Van der Lelie, Daniel 17
 Van Dien, Steve 28, 139
 Van Nostrand, Joy D. 89, 105
 Velappan, Nileena 37, 38, 39
 Venkatraman, Sankar 69
 VerBerkmoes, Nathan C. 20, 109, 111, 113, 124
 Victry, Kristin D. 69
 Viollier, P. 71
 Vuthoori, Madhavi 151

W

Wagner, Valerie 184
 Walbolt, Monica 40
 Walian, Peter J. 52, 53
 Walker, Chris B. 11, 28, 89, 96, 98
 Wall, Judy D. 28, 52, 58, 60, 89, 91, 92, 93, 95, 96, 97, 100, 103, 104
 Wang, Chunlin 84
 Wang, Q. 188
 Wang, Ting 63
 Wang, Xiaohu 162
 Wang, Yanbing 3
 Wang, Yu 133
 Wanger, Greg 10
 Ward, Mandy 158
 Ware, Doreen 203
 Watanabe, Masa 175
 Waterman, Michael S. 7
 Watson, David 25, 26, 105
 Weber, Peter K. 31, 32
 Weinberg, Michael V. 126
 Weiss, Shimon 160
 Werner, J. 71
 Whitaker, Rachel 20
 Wiback, Sharon J. 145
 Wiley, H. Steven 67, 69
 Wilmes, Paul 20, 111
 Witkowska, H. Ewa 52, 53
 Wong, Willy 185
 Wood, Stephan 44
 Woodard, Trevor L. 120
 Wozei, Eleanor 28, 89

Wu, Hongwei 149
 Wu, Liyou 25, 105
 Wu, W.-M. 26
 Wu, Weimin 105
 Wyrick, Judy M. 195

X

Xiong, Yijia 63
 Xu, Ying 149

Y

Yan, Bin 167
 Yan, Ping 63
 Yang, Chen 3
 Yang, Feng 117, 123
 Yang, Lee 52, 53
 Yang, Lin 46
 Yang, Yunfeng 162
 Yang, Zamin K. 89, 100, 107, 162
 Yannone, Steven M. 51
 Ye, Vincent 185
 Yeh, Yi Chun 83
 Yelton, A. Pepper 20
 Yen, Huei-Che 28, 89, 91, 92, 100
 Yilmaz, Ozlem 183
 Young, N.D. 15
 Yuan, Ling 133

Z

Zakharova, Natalia 61
 Zane, Grant 28, 89, 91
 Zemla, Adam 113
 Zhang, Haizhen 61
 Zhang, Peifen 137
 Zhang, Rui 128
 Zhang, Yang 152
 Zhao, Rui 128
 Zhong, Hui 46
 Zhou, Aifen 89, 91, 93
 Zhou, Jizhong 25, 26, 28, 89, 91, 92, 93, 95, 96, 100, 105, 162
 Zhuang, Kai 139
 Zhulin, Igor B. 155

Institutional Index

- American Type Culture Collection 12
- Argonne National Laboratory 4, 8, 22, 44, 62, 114, 116, 190
- Baylor College of Medicine 162
- Boston University 153, 154, 190
- Brookhaven National Laboratory 17, 46, 78, 206
- Brown University 164
- Burnham Institute for Medical Research 3, 116, 156
- California Institute of Technology 80
- Carnegie Institution 16, 137
- Carnegie Institution of Washington 146
- Carnegie Institute of Washington 205
- Carnegie Mellon University 80
- Case Western Reserve University 71
- Center for Advanced Research in Biotechnology 43
- Centocor 40
- Cold Spring Harbor Laboratory 154, 203
- Colorado School of Mines 146
- Columbus State University 130
- Cornell University 203, 209
- Dartmouth College 143
- Desert Research Institute 10
- Diversa Corporation 28, 34, 89, 100
- DOE Joint Genome Institute 10, 11, 14, 16, 25, 89
- East Carolina University 46
- Emory University 80
- Farasis Energy, Inc. 133
- Fellowship for Interpretation of Genomes 3, 156
- Florida State University 10
- Genomatica, Inc. 28, 139, 143, 145
- Georgia Institute of Technology 80
- Gibco, Invitrogen Corporation 132
- Hope College 134
- Howard Hughes Medical Institute 10, 89, 177, 179, 180
- Indiana University 10
- Institute for Information Transmission Problems 179
- Institute for Information Transmis-
- sion Problems 156
- Institute for Systems Biology 51, 151, 182
- Instituto de Biotecnología/UNAM 167
- Instituto de Tecnologia Quimica e Biologica 44
- Iowa State University 148, 209
- Johns Hopkins University 132, 158
- Joint Genome Institute 179
- Kansas State University 208
- Keck Graduate Institute 186
- Lawrence Berkeley National Laboratory 10, 11, 20, 25, 26, 28, 31, 47, 51, 52, 53, 56, 58, 60, 75, 83, 84, 85, 86, 89, 91, 92, 93, 95, 96, 97, 100, 101, 103, 105, 107, 177, 179, 180, 181, 185
- Lawrence Livermore National Laboratory 14, 20, 31, 32, 109, 111, 113, 175
- Los Alamos National Laboratory 14, 34, 37, 38, 39, 40
- Macrogenics Corporation 132
- Marine Biological Laboratory 8, 116
- Marine Biological Laboratory 190
- Massachusetts Institute of Technology 8, 10, 89, 92, 97, 116, 177, 180
- Miami University 25, 26, 28, 89, 91, 95
- Michigan State University 8, 12, 25, 130, 145, 162, 188
- Montana State University 66, 130
- National Institutes of Health 167
- National Renewable Energy Laboratory 146
- National Taiwan University 10
- North Carolina State University 207
- Oak Ridge National Laboratory 8, 20, 25, 26, 28, 34, 66, 67, 68, 69, 89, 92, 96, 100, 105, 107, 109, 111, 113, 116, 124, 155, 162, 174, 190, 195
- Oregon State University 135
- Pacific Northwest National Labora-
- tory 3, 4, 8, 10, 22, 61, 63, 66, 67, 68, 69, 114, 116, 117, 123, 128, 143, 145, 152, 156, 158, 162, 164, 190, 193
- Pomona College 142
- Portland State University 51
- Princeton University 10, 71, 73
- Purdue University 201
- Rensselaer Polytechnic Institute 164
- Requimte 44
- Rice University 157
- Samuel Roberts Noble Foundation 202
- Sandia National Laboratories 91, 101, 103
- Sandia National Laboratories, Livermore 89
- Scripps Research Institute 49, 51, 85, 126
- SoundVision Productions® 197, 199
- SRI International 137, 184
- Stanford University 14, 26, 71, 73, 75, 76, 83, 105, 123, 171, 185
- Temple University 89, 95, 100
- Texas A & M University 26, 105, 203
- The Institute for Genomic Research 167, 170
- The Wadsworth Center 164
- U.S. Geological Survey 20
- Universidad Nacional Autónoma de México 165
- University of California, Berkeley 10, 16, 20, 40, 52, 56, 58, 60, 89, 92, 97, 109, 111, 113, 177, 179, 180, 181
- University of California, Los Angeles 16, 160, 165
- University of California, Merced 175
- University of California, San Diego 75, 83, 122, 145, 146, 185
- University of California, San Francisco 40, 52, 53, 86
- University of Chicago 41
- University of Connecticut, Storrs 183
- University of Georgia, Athens 49,

- 51, 85, 126, 149, 209
 University of Kentucky, Lexington 133
 University of Massachusetts, Amherst 15, 35, 120, 122, 139, 141, 165, 167, 170
 University of Michigan, Ann Arbor 12
 University of Missouri, Columbia 28, 52, 58, 60, 89, 91, 92, 93, 95, 96, 97, 100, 103, 104
 University of Oklahoma, Norman 25, 26, 28, 89, 91, 92, 93, 95, 96, 100, 105, 162
 University of Puerto Rico, Mayaguez 26
 University of Southern California 3, 7, 8, 32, 142, 162, 190
 University of Tennessee, Knoxville 155
 University of Tennessee Health Science Center 167
 University of Toronto 139, 141, 167
 University of Washington, Seattle 11, 28, 89, 96, 98, 103, 104, 124, 186
 University of Waterloo 10
 University of Wisconsin, Milwaukee 152
 University of Wisconsin, Madison 176, 204
 Virtual Institute for Microbial Stress and Survival 10, 11, 25, 26, 28, 58, 89, 91, 92, 93, 95, 96, 97, 98, 100, 101, 103, 104, 105, 107, 177, 179, 180, 181
 Washington State University 61, 193
 Yale University 135

Notes

Notes

