DOE/SC-0140

Sections:
Systems Biology Strategies and Technologies for Understanding Microbes, Plants, and Communities
Structural Biology, Molecular Interactions, and Protein Complexes
Validation of Genome Sequence Annotation

U.S. DEPARTMENT OF ENERGY
Office of Science

Joint Meeting 2011

Genomic Science Awardee Meeting IX

and

USDA-DOE Plant Feedstock Genomics for Bioenergy Awardee Meeting

[Revised: April 14, 2011]

Crystal City, Virginia
April 10-13, 2011

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

http://genomicscience.energy.gov

Prepared by
Biological and Environmental Research 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

References

1. Brenner SE 1999 *Trends Genet*. 15 132-3
2. Galperin MY and Koonin EV 1998 *In Silico Biol*. 1 55-67
3. Jones CE, Brown AL and Baumann U 2007 *BMC Bioinformatics*. 8 170
4. Schnoes AM, Brown SD, Dodevski I and Babbit PC 2009 *PLoS Comput Biol.* 5 e1000605

This project has been funded with DOE grant BER KP 110201.

Structural Biology, Molecular Interactions, and Protein Complexes

240

Neutron Protein Crystallography Station User Facility

S. Zoe Fisher, Andrey Kovalevsky, Marat Mustyakimov, Marc-Michael Blum, Benno P. Schoenborn, Mary Jo Waltman, and **Paul Langan*** (Langan_paul@lanl.gov)

Bioscience Division, Los Alamos National Laboratory, Los Alamos, N.M.

Project Goals: PCS is a high-performance neutron beamline that forms the core of a BER-funded experimental User capability at Los Alamos Neutron Science Center (LANSCE) for investigating the structure and dynamics of proteins, biological polymers, and membranes.

Neutron diffraction is a powerful technique for locating hydrogen atoms, which can be hard to detect using X rays, and therefore can provide unique information about how biological macromolecules function and interact with each other and smaller molecules. This unique User capability is being used to investigate several enzymes that are important to USDA and DOE Genome Science program missions in renewable energy and the environment, with a view to understanding their detailed catalytic mechanisms. This new information is then being exploited to manipulate their performance and use. Neutron diffraction has also been crucial in revealing the structures and hydrogen bond arrangements in naturally occurring cellulose in lignocellulosic biomass and how they are rearranged by pretreatments to enhance conversion to biofuels. This information has led to the optimization of pretreatments to improve their cost-efficiency.

PCS Users have access to neutron beam time, deuteration facilities, protein expression and substrate synthesis with stable isotopes, a purification and crystallization laboratory, and software and support for data reduction and structure analysis. A HomeFlux X-ray system has been recently purchased that will allow users to collect X-ray data from the same samples used for neutron diffraction. The PCS beamline exploits the pulsed nature of spallation neutrons and a large electronic detector to efficiently collect wavelength-resolved Laue patterns using time-of-flight techniques. We encourage potential users to communicate with us before applying for beam time for technical guidance and help with proposal preparation.

For technical information about the PCS and experimental requirements contact Zoe Fisher (505) 665-4105 zfisher@lanl.gov or Paul Langan (505) 665 8125 langan_paul@lanl.gov

Proposal Submission: Proposals must be submitted using the process on the LANSCE website. To access the proposal submission site, go to the LANSCE home page, http://lansce.lanl.gov/. On this page click the tab "Lujan Center' and then the link 'Submit a Proposal.This will take you to the on-line submission system. Detailed instructions for preparing the proposals can be found on the proposal submission sites under "Step-by-Step Guide to Submitting an Online Proposal."

241

The Berkeley Synchrotron Infrared Structural Biology (BSISB) Program Overview

Hoi-Ying N. Holman[1,2]* (hyholman@lbl.gov), Hans A. Bechtel,[1,3] Rafael Gomez-Sjoberg,[1,4] Zhao Hao,[1,2] Ping Hu,[1,2] Michael C. Martin,[1,3] and Peter Nico[1,2]

[1]Berkeley Synchrotron Infrared Structural Biology Program, [2]Earth Sciences Division, [3]Advanced Light Sources, and [4]Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, Calif.

Project Goals: The Berkeley Synchrotron Infrared Structural Biology (BSISB) program is a national user facility for infrared spectromicroscopy and chemical microcharacterization of living cells.

The Berkeley Synchrotron Infrared Structural Biology (BSISB) program is a national user facility for infrared spectromicroscopy and chemical microcharacterization of biological systems. BSISB was initiated in 2010 to maintain a forefront research facility for infrared and optical characterization of chemistry in living cells with state-of-the-art instrumentation and expertise. The BSISB program has developed an integrated microfluidic synchrotron infrared (SIR) spectromicroscopy platform, which is a technique that is ideal for tracking the chemical composition and reactions in living cells during their adaptive responses to internal or external stimuli and perturbations. The BSISB program is also developing visible (VIS) hyperspectral/fluorescence microscopy approaches for simultaneously tracking changes in cellular morphology, structure, and other biological processes such as gene expression and signaling during SIR experiments. This new BSISB development of live-cell chemical biological imaging technologies will also be aided by a new generation of microfluidics platform. Our technological research and development effort will be accelerated by BSISB participating scientists with wide ranging research projects of bioenergy, medical, and environmental studies

that are important to DOE missions. Such synergistic interactions will provide interdisciplinary expertise and scientific critical mass to meet the emerging BSISB experimental challenges.

242

The National Center for X-Ray Tomography, a DOE Structural Biology Facility

Carolyn A. Larabell[1,2,3]* (CarolynLarabell@ucsf.edu) and Mark A. Le Gros[2,3]

[1]Department of Anatomy, University of California, San Francisco; and [2]Physical Biosciences Division, Lawrence Berkeley National Laboratory

http://ncxt.lbl.gov

Project Goals: DOE Structural Biology User Facility

The National Center for X-ray Tomography (NCXT) is a unique structural biology facility for high-resolution imaging of cells, biofilms and hydrated organic materials in general. The Center has developed soft X-ray tomography (SXT), which is a technique ideally suited to imaging subcellular architecture and organization in the native state. SXT is similar in concept to the well-established medical diagnostic technique computed axial tomography (CAT), except SXT is capable of imaging with a spatial resolution of 50 nm, or better. In SXT, cells are imaged using photons from a region of the spectrum known as the 'water window,' between the K shell absorption edges of carbon (284 eV, λ=4.4 nm) and oxygen (543eV, λ=2.3 nm). This results in quantitative, high-contrast images of intact, fully hydrated cells without the need to use contrast-enhancing agents. The method is particularly sensitive to the spatial distribution of carbon throughout microorganisms and provides information of great relevance to basic biological systems science, the production of biofuels and carbon sequestration. To image specific molecules with respect to cell structures, we have developed correlated imaging methods such as high numerical aperture cryogenic light microscopy. This multimodal approach allows fluorescently labeled molecules to be localized in the context of a high-resolution 3-D tomographic reconstruction of the cell. We will provide examples of data collected using SXT, which demonstrate that SXT is very well suited to DOE mission relevant research. The NCXT, as a DOE Structural Biology Facility, makes SXT available to national and international investigators.

243

Proteomic Scale Solution X-Ray Scattering and its Implications for Structural Biology

John Tainer and Greg Hura* (glhura@gmail.com)

Lawrence Berkeley National Laboratory

Project Goals: Ecosystems and Networks Integrated with Genes and Molecular Assemblies A multiscale systems approach to microbial bioremediation, carbon sequestration and energy production; from molecules to cells to communities.

High throughput solution structural analyses by small angle X-ray scattering efficiently enables the characterization of shape and assembly for nearly any purified protein. Crystallography has provided a deep and broad survey of macromolecular structure. Shape and assembly from SAXS in combination with available structures is often enough to answer critical mechanistic questions both enhancing the value of a structure and obviating larger crystallographic projects. Moreover, SAXS is a solution based technique, sample requirement are modest and compatible with many other biophysical methods. Here we present our high throughput SAXS data collection and analysis pipeline as applied to structural genomics targets, and metabolic pathways. Our goals of metabolic engineering and understanding protein mediated reactions rely on knowing the shape and assembly state of reactive complexes under an array of conditions. Given the number of gene products involved in metabolic networks, SAXS will play an important role in characterizing the structure of each individually, in complex with partners, and in various contexts. SAXS is well positioned to bridge the rapid output of bioinformatics and the relatively slow output of high resolution structural techniques.

243A‡

submitted post-press

CO_2 Capture by Amyloid Fibers

Dan Li*[1] (danli@mbi.ucla.edu), Hiroyasu Furukawa,[2] Omar M. Yaghi,[2] and **David Eisenberg**[1]

[1]UCLA-DOE, Institute for Genomics and Proteomics; [2]Department of Chemistry and Biochemistry, University of California-Los Angeles

Project Goals: To sequestrate CO_2 from flue gases by functionalized amyloid fibers.

Global warming accompanied with catastrophic consequences including food and water shortages, ecosystem irreversible change, and extreme weather events has generated great worries. A major cause of global warming is the increase of atmospheric greenhouse gas levels. 72% of the emitted greenhouse gases consist of carbon dioxide. Over the past 60 years, the global carbon emissions dramatically increased mainly from burning of carbon-based fossil fuels.

Capture of the CO_2 emitted from fossil fuel burning is of great importance for environmental protection.

Currently, aqueous monoethanolamine (MEA) is extensively applied to sequestrate CO_2 from flue gases. The amino group of MEA reacts with CO_2 forming carbamate. The reaction is rapid and requires no catalyst. However, MEA has many drawbacks. It is toxic, flammable and corrosive. MEA degrades in the presence of O_2 and CO_2 resulting in extensive amine loss and equipment corrosion as well as generating amine waste pollution. The regeneration of MEA demands high energy.

Here, we present a potential amine substitute—Lys-containing amyloid fibers. Amyloid fibers are protein aggregates with well-defined cross-β structures. As biological macromolecules, protein materials are biodegradable and environmental friendly. Protein amyloid fibers are repetitive structures. The hydrophobic core confers compactness and stability. Amyloid fibers cause a wide range of human pathologies, including Alzheimer's disease, dialysis-related amyloidosis and Parkinson disease. Attractively, the intrinsic structural properties of amyloid fibers make it possible to be functionalized as applicable materials. We have studied the CO_2 absorption capacity of amyloid fibers formed by Ac-VQIVYK-NH_2, an amyloidogenic sequence from the Alzheimer's tau protein. The dry fibers showed chemisorption of CO_2 by an isotherm experiment. In the VQIVYK fiber, the distance between lysines is ~4.8 Å, which may allow cooperativity of neighboring lysines. Therefore, compared to the monomeric peptide, lysine in fibers has a lower side chain pKa and is more favorable as the primary amine for carbamate formation.

Validation of Genome Sequence Annotation

244

NBC: The Naïve Bayes Classification Tool for Shotgun Metagenomics from Bacteria, Fungi, and Viruses

Gail Rosen[1]* (gailr@ece.drexel.edu), Aaron Rosenfeld,[1] Tze Yee Lim,[1] Yemin Lan,[1] and Christopher Blackwood[2]

[1]Drexel University (Electrical and Computer Engineering, Computer Science, Physics, and Biomedical Engineering Departments), Philadelphia, Penn.; and [2]Kent State University (Department of Biological Sciences), Kent, Ohio

Project Goals: We aim to produce a taxonomic classifier that runs quickly and can classify genomes (prokaryotic, eukaryotic, and viral). The naïve Bayes classification tool is implemented on a web site for public use, http://nbc.ece.drexel.edu. The database is soon to be expanded to include fungi and viruses, in order to expand its capabilities to soil and marine studies.

Datasets from high-throughput sequencing technologies have yielded a vast amount of data about organisms in environmental samples. Yet, it is still a challenge to assess the exact organism content in these samples because the task of taxonomic classification is too computationally complex to annotate all reads in a whole-genome shotgun dataset. An easy-to-use webserver is needed to process these reads. While many methods exist, only a few are publicly available on webservers, and out of those, most do not annotate all reads.

We introduce a webserver that implements the naïve Bayes classifier (NBC) to classify all metagenomic reads to their best taxonomic match. Results indicate that NBC can assign next-generation sequencing reads to their taxonomic classification and can find significant populations of genera that other classifiers may miss. We demonstrate that the tool can handle a complete pyrosequencing dataset in one day on a four-core server, and it gives the full lineage for each read, so that users can easily analyze the taxonomic composition of their datasets.

In addition to classification, we are working on developing a way to assess the novelty of sequences using detection theory. Using 5-fold cross-fold validation, we are able to achieve approximately 89% sensitivity and 95% specificity for species-level known/novel determination using the website top-hit likelihood scores. We are also working in conjunction with RDP to apply this technique to 16S rRNA sequences.

245

Assessing an Integrated Approach to Accurate Functional Annotation of Putative Enzymes in the Methanogen *Methanosarcina acetivorans*

Lucas Showman[1]* (lshowman@iastate.edu), Ethel Apolinario,[2] Libuse Brachova,[1] Yihong Chen,[3] Zvi Kelman,[3] Zhuo Li,[3] Kevin Sowers,[2] John Orban,[3] and **Basil J. Nikolau**[1]

[1]W. M. Keck Metabolomics Research Laboratory, Iowa State University, Ames; [2]Department of Marine Biotechnology, University of Maryland, Baltimore County; and [3]Institute for Bioscience and Biotechnology Research, University of Maryland College Park, Rockville

http://carb.umbi.umd.edu/g2f

Project Goals: We are using the methanogenic archaeon *Methanosarcina acetivorans* (MA) as a model organism to develop tools for rapid and reliable annotation and validation of protein function. The target genes are putative enzymes with detectable in vivo expression that have been selected after genome re-annotation.

The experimental approach begins with heterologous *E. coli* expression and purification of individual MA gene products. An initial ligand-binding screen of the purified protein using NMR spectroscopy determines whether candidate compounds can physically interact with the protein and act as putative substrates or products. Where possible, this is followed up with an experiment to see if the MA gene can complement an *E. coli* strain carrying a knock-out allele at the closest homolog.

In most if not all of our case studies, we find that the initial NMR screen is indicative of whether the function assignment is correct. Therefore this represents a method that may be suitable for accurate and efficient functional annotation of partially characterized enzymes without the need for developing protein-specific assays.

246

Functional Annotation of Putative Enzymes in *Methanosarcina acetivorans*

Ethel Apolinario,[1] Libuse Brachova,[3] Yihong Chen,[2] Zvi Kelman,[2] Zhuo Li,[2] Basil J. Nikolau,[3] Lucas Showman,[3] Kevin Sowers,[1] and **John Orban**[2,*] (jorban@umd.edu)

[1]Department of Marine Biotechnology, University of Maryland, Baltimore; [2]Institute for Bioscience and Biotechnology Research, University of Maryland College Park, Rockville; [3]W.M. Keck Metabolomics Research Laboratory, Iowa State University, Ames

Project Goals: The goal of the project is to develop rapid experimental approaches for accurate annotation of putative enzymatic functions. Targets of interest range from those with tentatively assigned function to hypotheticals.

Methane-producing organisms provide an efficient and cost-effective biofuel which is self-harvesting and can be distributed readily using infrastructure that is already in place. As with other genomes, however, accurate functional annotation in methanogens lags significantly behind the large body of sequence data, representing a sizable gap in our understanding of biology in these organisms. We are using the methanogenic archaeon, *Methanosarcina acetivorans* (MA), as a model system for developing experimental tools for rapid and reliable annotation and validation of function. The target genes are putative enzymes in MA with detectable *in vivo* expression.

Our experimental approach utilizes a combination of methods for rapid function assignment. NMR spectroscopy is used to screen for putative substrates, products, or their structural analogs. Where possible, we have followed up on function assignments by checking to see if the MA gene can complement the corresponding *E. coli* knockout. We have used this approach to both validate and correct functional assignments in MA target genes, as will be illustrated with examples. Further, insights into the functional annotation of "hypotheticals" are being obtained by integrating mass spectrometry based metabolite profiles of gene knockouts with NMR-based approaches and these will also be discussed.

247

The Ribosomal Database Project: Tools and Sequences for rRNA Analysis

J.R. Cole* (colej@msu.edu), Q. Wang, B. Chai, J.A. Fish, D.M. McGarrell, and **J.M. Tiedje**

Center for Microbial Ecology, Michigan State University, East Lansing

Project Goals: The Ribosomal Database Project (RDP; http://rdp.cme.msu.edu) offers aligned and annotated rRNA sequence data and analysis services to the research community. These services help researchers with the discovery and characterization of microbes important to bioenergy production, biogeochemical cycles, greenhouse gas production, and bioremediation.

Our view of the evolutionary relationships among life forms on Earth has been revolutionized by the comparative analysis of ribosomal RNA sequences. Life is now viewed as belonging to one of three primary lines of evolutionary descent: Archaea, Bacteria and Eucarya. This shift in paradigm has not only challenged our understanding of life's origin, but also provided an intellectual framework for studying extant life—particularly the vast diversity of microorganisms. Ribosomal RNA diversity analysis using genes amplified directly from mixed DNA extracted from environments has demonstrated that the well-studied microbes described by classical microbial systematics represent only a small percentage of diversity. The use of rRNA to explore uncharacterized diversity had become such a relied-upon methodology that by 2008, 77% of all INSDC bacterial DNA sequence submissions described an rRNA sequence, and only 2% of these entries had a Latin name attached (valid or otherwise; Christen, 2008)! Examining the RDP's collection of quality rRNA sequences demonstrates that cultivated organisms represent only a fraction of observed rRNA diversity, and currently available genome sequences cover an even smaller slice of this cultivated fraction (Fig. 1). Phylogenetically informed selection of sequencing candidates, as done in the GEBA Project, can help improve genome coverage of diversity represented by cultivated organisms (Wu et al., 2009), and single cell sequencing can provide partial genome data for uncultivated organisms; but it will be years before these techniques are able to make practical progress towards tackling the immense diversity represented by the collection of rRNA sequences. In fact, it is our knowledge of rRNA diversity that is guiding these efforts.

In the current release (January 2011), RDP offers 1,498,677 aligned and annotated quality-controlled public bacterial and archaeal rRNA sequences along with tools that allow researchers to examine their own sequences in the context of the public sequences (Cole et al., 2009). In addition, 8,770 researchers have over 5.9 million private pre-publication

sequences in the *my*RDP account system. On average, the RDP website is visited by over 8,400 researchers (unique ip addresses) from more than 150 countries in over 21,000 analysis sessions each month, while Web Services (SOAP) interfaces to these RDP analysis functions process an additional 8.6 billion bases of sequence per month for high-volume users running their own analysis queues. In addition, since its release in May 2008, the RDP Pyro Pipeline has been used by over 1,800 researchers (unique e-mail addresses) to process their own high-throughput current-generation rRNA sequence data. In addition, the command-line version of the RDP Classifier, designed for users needing local high-throughput analysis, has been downloaded over 3,400 times since its release on SourceForge (Wang et al., 2007). The RDP has also been working with standards bodies, such as the Genomic Standards Consortium (GSC; http://gensc.org/gsc/) and the Terragenome Consortium (http://www.terragenome.org) to help define environmental annotation standards for rRNA and other environmental marker gene libraries, and to assure that RDP is ready for the new standards. These results have been incorporated into the new MIMARKS (Minimal Information about a MARKer gene Sequence; Nature Biotechnology, in press) covering rRNA and environmental gene sequences.

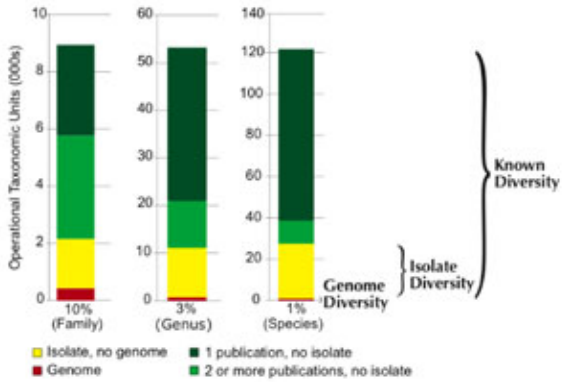

Figure 1. Clusters from a total of 668,513 high-quality near-full-length bacterial rRNA sequences, including sequences from all 2,179 bacterial genomic sequences available from GenBank RefSeq on 23 September 2010 were clustered using the RDP mcClust tool implementing a memory constrained complete-linkage algorithm. Cluster distances approximate the given taxonomic rank.

The RDP is supported by the Office of Science (BER), U.S. Department of Energy, Grant No. DE-FG02-99ER62848

References

1. Christen, R. 2008. Global sequencing: a review of current molecular data and new methods available to assess microbial diversity. *Microbes and Environments* 23:253-268.
2. Cole, J. R., Q. Wang, E. Cardenas, J. Fish, B. Chai, R. J. Farris, A. S. Kulam-Syed-Mohideen, D. M. McGarrell, T. Marsh, G. M. Garrity, and J. M. Tiedje. 2009. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.* 37:D141-D145
3. Wang, Q., G. M. Garrity, J. M. Tiedje, and J. R. Cole. 2007. Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol.* 73:5261-5267.
4. Wu, D., P. Hugenholtz, K. Mavromatis, R. Pukall, E. Dalin, et al. 2009. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* 462:1056-1060.