Technology Development

Imaging

77

Electron Tomography of Intact Microbes

Kenneth H. Downing (khdowning@lbl.gov)

Lawrence Berkeley National Laboratory, Berkeley, CA

Electron tomography is developing as an effective tool to study subcellular structure at the molecular level. While the thickness of samples that can be studied is limited to a fraction of a micrometer, the resolution is, in principal, sufficient to identify many of the major macromolecular complexes and thus gain insights on their location and interaction. Such information will be essential for the ultimate goals of understanding and building complete computational models of the microbes.

In our initial work to develop electron tomography of intact cells and explore its limits of applicability, we have established culture and preparation conditions for a number of small microbes that may be potential targets for this work. The thinner cells are somewhat better suited for study in whole-cell preparations, but we have shown that we can record 2-D projection images by electron microscopy of each of these in frozen-hydrated preparations showing substantial internal detail. We thus retain the native state with no stain or other contrast enhancements, but can see a wealth of internal structures. Frozen-hydrated samples, though, are difficult to work with for several reasons. Aside from the technical issues of specimen preparation and data recording, which are now handled quite routinely, the dense molecular packing and high protein density within bacterial cells makes interpretation difficult for many of the cell components. Large, extended structures, such as cytoskeletal filaments and condensed nucleic acids should be fairly easily discriminated, and once the target resolution range is achieved we expect to be able to identify the major protein complexes.

In the meantime, we have been using a more conventional approach of examining sections of embedded samples. Specimens prepared by high pressure freezing and freeze substitution provide quite good preservation. For example, in eukaryotic samples microtubules provide measure of resolution, and the 40-Angstrom thick protofilaments are often well resolved. This approach is being used to investigate chromium sequestration in *Arthrobacter oxydans* and morphological changes following stress in *Desulfovibrio*, in collaboration with Hoi-Ying Holman at LBNL.

In both frozen-hydrated and embedded samples, we need to develop the ability to identify specific molecular components by labeling. The equivalent of GFP for electron microscopy is a goal of several groups, and several promising approaches are being followed. As a first step, one can use heavy metal cluster labeling of antibodies in the manner as fluorescent antibodies are used at the light microscope level.

Procedures for data collection and processing that overcome the main bottlenecks in generating 3-D representations of the cells have been developing rapidly over the

last few years. Several options now exist for software to control the electron microscope for data collection. To a greater or lesser extent, these relieve the tedium of recording the large number of images required for tomography and the large number of steps needed in collecting each image, thus enabling much faster data collection and ultimately far higher quality data. There is still work that needs to be done to improve the data collection stage, but the remaining rate limiting steps have more to do with visualization of the large data volumes and automated searches through the volume to identify components of interest.

78

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

Gang Bao (gang.bao@bme.gatech.edu)

Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA

To reach the goals of the DOE Genomics:GTL program, there is an urgent need to study individual proteins and multi-protein complexes in microbes. Currently, there is a lack of novel labeling reagents for performing protein intracellular localization and mapping studies. There are few tools that can be used to identify individual proteins and characterize multi-protein complexes in microbial cells, and to visualize and track assembly and disassembly of multi-protein molecular machines. There are no methods to study simultaneous co-localization and dynamics of different intra-cellular processes with high spatial resolution. To meet this challenge, in this study we propose 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. This innovative molecular imaging approach integrates peptide-based cellular delivery, protein targeting/tagging, light microscopy and electron microscopy. Specifically, we propose to develop multifunctional quantum-dot bioconjugates consisting of (1) a quantum dot of 2-6 nm in size encapsulated in a phospholipid micelle, (2) delivery peptides and protein targeting ligands (called adaptors) conjugated to the surface of the QD through a biocompatible polymer. 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 will be 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 will then be 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 of protein detection and localization.

A highly interdisciplinary team has been assembled for this DOE project, with participating faculty members from four universities (Georgia Tech, Emory U, Carnegie Mellon U, and Caltech). The long-term goal is to develop a new multifunctional nanoparticle based molecular imaging platform with enhanced sensitivity, specificity, and spatial resolution. During the proposed three-year period, we will specifically: (1) design, synthesize and characterize quantum dots (QDs) with controlled properties and surface modifications for conjugation with ligands and peptides; (2) conjugate specific ligands such as antibodies and small organic molecules to QDs; (3) develop a peptide-based approach for delivering nanoparticle bioconjugates into

microbial cells; (4) perform fluorescence imaging and electron microscopy to identify, localize, and track proteins and protein complexes. This platform technology will have a wide range of biological and biomedical applications relevant to the Genomes to Life program at DOE, including an improved understanding of multi-protein molecular machines, protein assemblies/networks, and detailed protein functions.

79

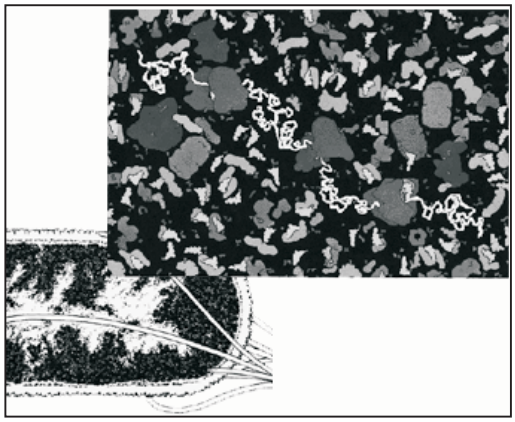
Single Molecule Imaging of Macromolecular Dynamics in a Cell

Jamie H. D. Cate[1-3] (jcate@lbl.gov), Jennifer Blough[1], Hauyee Chang[1], Raj Pai[2], Abbas Rizvi[1], Chung M. Wong[1], Wen Zhou[1], and Haw Yang[1,3] (hawyang@uclink.berkeley.edu)

[1]Department of Chemistry and [2]Department of Molecular and Cell Biology, University of California, Berkeley, CA; and [3]Lawrence Berkeley National Laboratory, Berkeley, CA

We are developing technologies and strategies to image individual proteins and multi-protein complexes in microbes, in order to provide high-resolution and quantitative information on the function of macromolecules in the context of the cell.

The interior of a cell is densely packed with macromolecules. This picture is perhaps best illustrated by a drawing by David Goodsell displayed on the left, showing a number of ribosomal complexes along a RNA chain in the crowded interior of a bacterium[1]. The cellular interior is filled with confined protein molecules and supramolecular complexes that are likely to exhibit different thermal structural fluctuations *in vivo* than those seen *in vitro*[2]. For instance, the diffusion constant of green fluorescent protein (GFP) has been found to decrease four-fold to ten-fold inside a cell relative to diffusion in water[3]. How, then, do the dynamics of biomolecules in crowded environments affect chemical processes in the cell? How do the rates of enzymatic reactions measured *in vivo* compare with those measured *in vitro*?



These questions are very difficult to address by ensemble-averaged assays. A biomolecule inside a cell, constrained in its diffusion, may show a broad location-dependent distribution in its dynamical properties that are distinctly different from those measured *in vitro*. The convoluted spatio-temporal dynamics in cells make it very hard to quantitatively study the various molecular dynamics of a functioning biomolecule. We anticipate that single-molecule spectroscopy, due to its capability of obtaining the individual dynamics from a distribution, will prove invaluable in efforts to unravel how microscopic, molecular interactions impact macroscopic biological functions.

In order to measure macromolecular function and dynamics in the cell, we are developing a single-molecule spectrometer with 3D single-particle tracking capabili-

ties. In an experiment, the biomolecule to be tracked will be conjugated to a tracer element, a surface-passivated nanoparticle that reports the precise location of the biomolecule. In addition to the tracer element, the tracked biomolecule will be labeled with fluorescent probes at strategic sites to allow for simultaneous studies of macromolecular dynamics such as conformational rearrangements, and association and dissociation of macromolecular complexes. We are studying the protein synthesis machinery in *Deinococcus radiodurans* as a model system with which to develop these technologies. Our goal is to establish single-molecule spectroscopy as a general approach to study macromolecules in living organisms.

References

1. Goodsell, D.S. The Machinery of Life (Springer-Verlag, New York, 1998).

2. Ellis, R.J. Macromolecular crowding: obvious but underappreciated. *Trends Biochem Sci* **26**, 597-604 (2001).

3. Elowitz, M.B., Surette, M.G., Wolf, P.E., Stock, J.B. & Leibler, S. Protein mobility in the cytoplasm of Escherichia coli. *Journal of Bacteriology* **181**, 197-203 (1999).

80

Developing a Hybrid Electron Cryo-Tomography Scheme for High Throughput Protein Mapping in Whole Bacteria

Huilin Li (hli@bnl.gov) and James Hainfeld

Biology Department, Brookhaven National Laboratory, Upton, NY

The structures of biological molecular assemblies and their locations inside cells are keys to understanding their functions. Fluorescence microscopy in combination with phase contrast light microscopy is successful in protein localization, but it is limited by its low resolution. This is a serious problem in studying smaller cells such as bacteria with a size of only 1 micrometer. Electron cryo-tomography is an alternative approach to this problem. It provides close to native structure preservation and significantly higher resolution (in the range of 5 to 10nm) three-dimensional structures. However because of the particularly crowded bacterial cellular environment, it is currently difficult to unambiguously identify most proteins. We are developing a hybrid approach, by taking advantage of ultra-structural visualization capability of the cryo-electron microscopy (cryo-EM) and the heavy metal cluster label detection capability of the scanning transmission electron microscopy (STEM) to achieve simultaneously three-dimensional structural visualization and protein mapping. Toward this goal, we will first develop an optimum procedure to label microbial cells, while keeping their structures minimally disturbed. A novel in situ bi-modal tomography protocol of cryo-EM and cryo-STEM will also be developed. To make this method a high throughput tool, universal labels targeted to genetically encoded signatures, such as the Ni-NTA-gold label and 6X-histidine tag system will be developed. The technique will be applied initially to mapping and visualization of the bacterial "cytoskeleton" system and heavy metal resistance protein complexes in *Ralstonia metallidurans*, a microbe of direct DOE interest.

The goals of the project are:

1. To develop a hybrid electron cryo-tomography scheme. We will develop an *in situ* cryo-TEM and cryo-STEM tomography bimodal imaging scheme. The two

tomograms from TEM and STEM tilt series of bacterium embedded in amorphous ice are merged to achieve a simultaneous mapping and visualization of protein complexes in bacteria.

2. To develop a high throughput bacterial labeling strategy based on a 3nm gold particle and genetically encoded signature labeling system, such as Ni-NTA-gold and 6X His-tag proteins. A procedure for mild cell fixing and permeabilization will be developed to allow for label access to the inside of the cell. The procedure is based on the established bacterial cell treatment method in immuno-fluorescence microscopy. After labeling, the bacterial cells will be rapidly frozen in vitreous ice for imaging.

3. To apply the developed methods for high-resolution mapping and visualization of the "cytoskeleton" proteins and multiple heavy metal resistance complexes in *Ralstonia metalliduran*.

81

Probing Gene Expression in Living Bacterial Cells One Molecule at a Time

X. Sunney Xie[1] (xie@chemistry.harvard.edu), Jie Xiao[1], Long Cai[1], and Joseph S. Markson[2]

[1]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA and [2]Department of Chemistry, University of Cambridge, Cambridge, U.K.

We demonstrate for the first time continuous real-time monitoring of gene expression in individual living cells with single-copy sensitivity for a protein. Our approach is based on a modified reporter protein, a short-lived beta-galactosidase (beta-gal), which hydrolyzes a fluorogenic and membrane-permeable substrate and degrades inside a cell within a few minutes. The enzymatic amplification of the hydrolysis product allows us to observe fluorescence bursts corresponding to stochastic expression of the reporter protein in a single *E. coli* cell. Each burst is triggered by the dissociation of the *lac* repressor from the *lac* operator on the *E. coli* chromosome. Moreover, the time traces of the fluorescence bursts exhibit quantized levels corresponding to beta-gal molecules generated and degraded one at a time. The implication of this work to gene expression profiling as well as system-wide studies of gene regulation will be discussed.

Protein Production and Molecular Tags

82

Developing a High Throughput Lox Based Recombinatorial Cloning System

Robert Siegel[1], Nileena Velappan[2], Peter Pavlik[2], Leslie Chasteen[2], **Andrew Bradbury**[2] (amb@lanl.gov)

[1]Pacific Northwest National Laboratory, Richland, WA and [2]Los Alamos National Laboratory, Los Alamos, NM

The selection of affinity reagents (antibodies, single chain Fvs - scFvs) against protein targets can be done using a number of different systems, including phage, phagemid, bacterial or yeast display vectors. Genetic selection methods have also been developed based on yeast two hybrid and enzyme complementation systems. In general, selection vectors are not suitable for subsequent scFv production. Furthermore, once scFvs have been selected, they can be usefully modified by cloning into other destination vectors (e.g. by adding dimerization domains, detection domains, eukaryotic expression in eukaryotic vectors etc.). However, this is relatively time consuming, and requires checking of each individual construct after cloning. An alternative to cloning involves the use of recombination signals to shuttle scFvs from one vector to another. These have the advantage that DNA restriction and purification can be avoided. Such systems have been commercialized in two general systems: Gateway™, uses lambda att based recombination signals, while Echo™ uses a single lox based system to integrate a source plasmid completely into a host plasmid.

We have examined the potential for using heterologous lox sites and cre recombinase for this purpose. Five apparently heterologous lox sites (wild type, 511, 2372, 5171 and fas) have been described. A GFP/lacZ based assay to determine which of these were able to recombine with each other was designed and implemented. Of the five, three (2372, 511 and wt) were identified which recombined with one another at levels less than 2%.

To use recombination as a cloning system, it is important to be able to select against host vectors which do not contain the insert of interest. Two toxic genes were examined for this purpose. The tetracycline gene confers sensitivity to nickel, while the sacB gene confers sensitivity to sucrose. We confirmed these sensitivities, although found that some antibiotic resistances interfere with survival of bacteria hosting non-tetracycline containing plasmids.

In preliminary experiments we have demonstrated that recombination from one plasmid to another, using 2372 and wild type lox sites and sacB or tetracycline, can occur in vivo at very high efficiency. This opens the possibility of using this system to easily transfer scFvs after selection to other plasmids. However, the utility of this system is not limited to scFvs - any DNA fragment (gene, open reading frame, promoter etc.) can easily be shuttled from one plasmid to another using these lox based signals.

Antibody libraries have been made using these lox sites, and are in the process of being evaluated.

83

Methods for Efficient Production of Proteins and High-Affinity Aptamer Probes

Michael Murphy, Paul Richardson, and **Sharon A. Doyle**

DOE Joint Genome Institute, Walnut Creek, CA

With genome sequencing efforts producing vast amounts of data, attention is now turning towards unraveling the complexities encoded in the genome: the protein products and the cis-regulatory sequences that govern their expression. Understanding the spatial and temporal patterns of protein expression as well as their functional characteristics on a genomic scale will foster a better understanding of biological processes from protein pathways to development at a systems level. Presently, the main bottlenecks in many proteomics initiatives, such as the development of protein microarrays, remain the production of sufficient quantities of purified protein and affinity molecules or probes that specifically recognize them. Methods that facilitate the production of proteins and high affinity probes in a high-throughput manner are vital to the success of these initiatives. We have developed a system for high-throughput subcloning, protein expression and purification that is simple, fast and inexpensive. We utilized ligation-independent cloning with a custom-designed vector and developed an expression screen to test multiple parameters for optimal protein production in *E. coli*. A 96-well format purification protocol was also developed that produced microgram quantities of pure protein. These proteins were used to optimize SELEX (Systematic Evolution of Ligands by Exponential Enrichment) protocols that use a library of DNA oligonucleotides containing a degenerate 40mer sequence to identify a single stranded DNA molecules (aptamers) that bind their target protein specifically and with high affinity (low nanomolar range). Aptamers offer advantages over traditional antibody-based affinity molecules in their ease of production, regeneration, and stability, largely due to the chemical properties of DNA versus proteins. These aptamers were characterized by surface plasmon resonance (SPR) and were shown to be useful in a number of assays, such as western blots, enzyme-linked assays, and affinity purification of native proteins.

This work was performed under the auspices of the U.S. Department of Energy, Office of Biological and Environmental Research, by the University of California, under Contracts No. W-7405-Eng-48, No. DE-AC03-76SF00098, and No. W-7405-ENG-36.

84

Development of Multipurpose Tags and Affinity Reagents for Rapid Isolation and Visualization of Protein Complexes

M. Uljana Mayer, Liang Shi, Yuri A. Gorby, David F. Lowry, David A. Dixon, Joel G. Pounds, and **Thomas C. Squier** (Thomas.Squier@pnl.gov)

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

Our long-term goal is to develop high-throughput methods for the rapid and quantitative characterization of protein complexes in microbial cells *in vivo*. The initial focus of the proposal will be on *S. oneidensis* MR-1, whose metabolism is important in understanding both microbial energy production and environmental remediation. However, these strategies will be applicable to a wide range of microorganisms and will permit the identification of environmental conditions that affect the expression of critical proteins required for the formation of adaptive protein complexes that facilitate bacterial growth. Our hypothesis is that identifying dynamic changes in these adaptive protein complexes will provide important insights into the metabolic regulatory strategies used by these organisms to adapt to environmental changes.

We propose to implement a strategy focusing on the development of multiuse protein tags engineered around a tetracysteine motif (i.e., CCXXCC), which has previously been shown to provide a highly selective binding site for cell permeable arsenic-containing affinity reagents that can be used to first identify and then validate protein complexes in living cells (*Griffin et al., 1998; Adams et al., 2002*). Taking advantage of the large increase in the fluorescence signal associated with binding the proposed fluorescent affinity reagents to the protein tag, it will be possible to use on-line detection to monitor affinity isolation of protein complexes and rapidly identify the proteins in the complex using mass spectrometry. Identification of low-affinity binding interactions in protein complexes is possible by engineering protein crosslinkers onto the bisarsenical affinity reagents. Furthermore, these same protein tags and affinity reagents will permit real-time visualization of steady-state protein abundance and protein-protein interactions, permitting validation of identified protein complexes under cellular conditions and the high-throughput identification of metabolic flow through defined biochemical pathways in response to environmental conditions. Ultimately, these methods will permit an optimization of useful metabolic pathways to fulfill Department of Energy (DOE) goals involving efficient energy utilization, carbon sequestration, and environmental remediation. To accomplish these goals, we propose three specific aims: (1) Identify multipurpose tags with optimized sequences for differential labeling using cell permeable orthogonal fluorescent probes, (2) Optimize expression and *in vivo* labeling of tagged proteins in *S. oneidensis MR-1*, and (3) Develop improved affinity reagents with optional photocrosslinking extensions for stabilizing and identifying cellular protein complexes.

In the next three years we will proceed to fulfill the following three aims:

Aim 1: *Identify Multipurpose Tags with Optimized Sequences for Differential Labeling Using Cell Permeable Orthogonal Fluorescent Tags*. We propose to develop cell permeable reagents that selectively associate with unique tags engineered into bacterial proteins which, in turn, permit highly specific affinity purification strategies for the isolation of protein complexes in *Shewanella oneidensis* and other microbes. These methods are based upon initial results using the fluorescent reagent FlAsH

(Fluorescein Arsenical Helix binder), which specifically interacts with tags containing tetracysteine motifs (i.e., CCXXCC). Optimization of the structure of tags for specific affinity reagents will be accomplished using peptide libraries and computational methods. The structure of the affinity reagents bound to the peptide tags will be determined, allowing for the rational redesign of these affinity reagents to enhance their binding specificities for new affinity reagents.

Aim 2: *Expression and in vivo Labeling of Tagged Proteins in Shewanella oneidensis*. Targeted genes will be cloned under their own promoter into shuttle vectors for *in vivo* expression of tagged proteins following identification of candidate proteins that are expected to form complexes. Expression in *S. oneidensis* MR-1 will be optimized. Tags will be developed to have a minimal impact on cellular metabolism, as determined by measuring their effect on maximal growth rate and molar growth yield in wild-type and modified organisms. Expressed proteins will be identified using fluorescent affinity reagents that recognize specific binding motifs on tagged proteins, permitting the rapid characterization of rates of protein expression and turnover under defined environmental conditions. Optimization of affinity reagents for *in vivo* labeling will involve expression of tagged Aequorea-derived fluorescent proteins (AFP) proteins, whose extent of modification will be measured using fluorescence resonance energy transfer (FRET) methods.

Aim 3: *Develop Improved Affinity Reagents with Optional Photocrosslinking Extensions for in vivo Stabilization and Identification of Protein Complexes*. Purification will involve immobilization of affinity reagents on solid supports. Purified proteins will permit a quantitative characterization of affinity and specificity of affinity labels. Complex purification using both protein encoded tags and bisarsenical probe-based affinity reagents will be optimized. Building upon the scaffolding of known cell-permeable reagents, we propose to develop cleavable crosslinking reagents that stabilize protein complexes and facilitate their isolation and identification using mass spectrometry. Thus, transient protein associations, such as those responsible for fast metabolic control mechanisms, may be identified. Following adduct formation between affinity reagents and protein tags, light activation of the photoreactive moieties will permit crosslinking to binding partners. Following isolation of protein complexes and trypsin digestion, crosslinked peptides will be isolated using an engineered affinity tag, and identified using mass spectrometry.

References

Adams, S.R., R.E. Campbell, L.A. Gross, B.R. Martin, G.K. Walkup, Y. Yao, J. Llopis, and R.Y. Tsien (2002) *New bisarsenic Ligands and tetracysteine motifs for protein labeling in vitro and in vivo: Synthesis and biological applications.* J. Am. Chem. Soc. **124**: 6063-6076.

Griffin, B.A., S.R. Adams, and R.Y. Tsien (1998) *Specific covalent labeling of recombinant protein molecules inside live cells*. Science **281**: 269-272.

85

Development of Genome-Scale Expression Methods

Frank Collart[1] (fcollart@anl.gov), Gerald W. Becker[2], Brian Hollaway[2], Yuri Londer[1], Marianne Schiffer[1], and Fred Stevens[1]

[1]Argonne National Laboratory, Argonne, IL; and [2]Roche Protein Expression Group, Indianapolis, IN

The capability to express proteins in heterologous systems has been an important enabling feature for structural and functional studies of proteins. Although, recent advances in expression technology have significantly increased our capability for the expression of microbial proteins, a significant fraction of proteins encoded by the genome still cannot be expressed in a usable form. We will address these challenging expression problems by application of novel cellular and cell-free technologies to optimize the expression of "insoluble" cytoplasmic and periplasmic proteins. As part of this process, we will evaluate domain-based cloning and expression methods for high molecular weight proteins and putative soluble domains of membrane proteins. The domain-based approach provides an alternative to full length expression and is often used in traditional benchtop approaches. Application of this approach will allow production of soluble domains for many proteins and enable biophysical and biochemical characterization and affinity tag production. These protein resources will support the GTL program and the information gained from these domains may ultimately be used to design a successful strategy for production of the full length protein. Proposed goals include:

1. Extending the capabilities of present high throughput expression platforms to address challenging areas for expression:
   - Apparent insoluble cytoplasmic proteins
   - High molecular weight proteins
   - Soluble domains of membrane proteins
   - Periplasmic proteins
2. Generation of a database using experimental and historical expression data to facilitate development of predictive methods for optimization of expression strategy.
3. Promoting interaction with GTL collaborators to prioritize experimental workflow and facilitate distribution of research resources.

Although automation and high throughput methods can ameliorate some of the cost of protein production, comprehensive protein production strategies will require a balance between optimization of automated methods to enable the cost effective production of clones and proteins and the development of more complex expression strategies for difficult proteins. A major focus of this project is the extension of traditional plate-based methods to address challenging expression problems for proteins from *Shewanella oneidensis* and *Geobacter sulfurreducens*. This dual strategy leverages the cost effectiveness of HT methods to conserve resources and focus on the significant fraction of cellular proteins that remain difficult expression problems but are essential to the undertaking of a system biology approach in understanding microbial cells.

86

Chemical Methods for the Production of Proteins

Stephen Kent (skent@uchicago.edu)
University of Chicago, Chicago, IL

Background

There is a critical need for methods of producing proteins whose existence is predicted by bioinformatic analysis of microbial genome sequence data, in order to undertake their biophysical and functional characterization. Powerful recombinant DNA-based methods exist for the production of proteins in genetically engineered microorganisms or in cell-free translation systems. However, small {<80 amino acid} proteins (~15% of a typical genome) and integral membrane proteins (~25% of a typical genome) have so far proved to be refractory to ready production by these methods. We will prototype novel methods for the high throughput production of milligram amounts of these special classes of proteins using chemical synthesis ['Synthesis of native proteins by chemical ligation.' Dawson, P.E., Kent S.B.H., *Ann. Rev. Biochem*., **69**, 925-962 (2000)].

Our goal is to address the known limitations of *chemical* protein synthesis, based on our intimate understanding of the current state of the art, as exemplified by the total chemical synthesis of the model protein crambin ['Total chemical synthesis of crambin.' Bang, D., Chopra, N., Kent, S.B.H. *J. Am. Chem. Soc.*, In press; 'A one-pot total synthesis of crambin.' Bang, D., Kent, S.B.H., *Angewandte Chemie*, submitted].

Technology Development

The first phase of our research program is focused on the development and optimization of methods aimed at filling the gaps in the tool kit of chemical protein synthesis techniques. These include:

A. Chemical synthesis of peptide-thioesters
   i. Nucleophile-stable thioester-generating resins for solid phase synthesis
   ii. Activation and coupling in the absence of base
   iii. Flow deprotection and cleavage
   iv. High throughput verification of amino acid sequences

B. Ligation at non-cysteine residues
   i. 'Pseudo' native chemical ligation
   ii. Extended native chemical ligation
   iii. 'Traceless' chemical ligation

C. Polymer-supported chemical ligation (solid phase protein synthesis)
   i. Linker chemistries
   ii. Analytical control

In this way, we will develop a practical chemical protein synthesis technology applicable to the rapid preparation of multiple milligram amounts of small and integral membrane protein targets based on predicted gene sequence data. We will then prototype the application of these methods to selected proteins of the model organism *Shewanella oneidensis*.

Significance & Impact

Using the chemical ligation approach the science of chemistry can now be applied, without limitation, to the study of the protein molecule. Chemical synthesis enables the application of all the ingenuity of the modern chemical methods to be applied to the study of the molecular basis of protein function. Applications range from the straightforward replacement of individual amino acid building blocks to much more elaborate and ingenious chemical schemes to engineer new forms of the protein molecule:

- Non-coded amino acids can be incorporated without limitation as to kind, position within the polypeptide chain, and number of substitutions. Non-amino acid building blocks can also be used. For example, a bicyclic β-turn mimetic of fixed geometry was introduced into the HIV-1 protease molecule.[1]
- Post-translational modifications: glycoproteins[2] and glycoprotein mimetics.
- Chemical synthesis can be used to introduce nmr probe nuclei at specific single atom sites in a protein molecule, in any desired number and combination. This can be invaluable for sorting out residue assignments in overlapping regions of the spectra†. Using expressed protein ligation, it is readily possible to mix and match biosynthetically isotope-enriched domains with unlabelled domains in order to simplify the interpretation of nmr spectra of larger proteins.[3]
- Reporter moieties for physical techniques such as EPR or fluorescence spectroscopy can be introduced at will at any desired location within the protein molecule being studied.

Radical re-engineering of the protein molecule has included: building in chemical cleavage sites to unzip the peptide chain at will for protein footprinting[4]; the preparation of proteins containing cyclic polypeptide chains[5]; the construction of topological analogues of proteins (e.g. two N-terminals, no C-terminus; interpenetrating cyclic polypeptide chains[6]).

†This will have particular application to polytopic helical integral membrane proteins; these molecules contain large numbers of identical hydrophobic amino acids in similar chemical environments. Labeling subsets of these residues with nmr probe nuclei will be essential to interpretation of high resolution magic angle spinning nmr spectra of membrane protein preparations.

References

1. Baca M., Alewood P., Kent S.B.H., *Protein Science,* **2**, 1085-1091 (1993).
2. Marcaurelle LA, Mizoue LS, Wilken J, Oldham L, Kent SB, Handel TM, Bertozzi CR, *Chemistry*. **7**, 1129-32 (2001)
3. Cowburn D, Muir TW, *Methods Enzymol.* **339**, 41-54 (2001).
4. Tom W. Muir, Philip E. Dawson, Michael C. Fitzgerald, Stephen B.H. Kent. *Chemistry & Biology,* **3**, 817-825 (1996).
5. Craik DJ, Simonsen S, Daly NL, *Curr Opin Drug Discov Devel.* **5**, 251-60 (2002).
6. Blankenship JW, Dawson PE, *J Mol Biol*., **327**, 537-548 (2003).

87

A Combined Informatics and Experimental Strategy for Improving Protein Expression

John Moult (jmoult@tunc.org), Osnat Herzberg, Frederick Schwarz, and Harold Smith

Center for Advanced Research in Biotechnology, Rockville, MD

The impetus for this project arose out of experience with microbial protein expression in a structural genomics project. We have explored the expression of over 300 non-membrane proteins from *Haemophilus influenzae* and *E. coli* using state of the art over-expression protocols. Our findings are similar to those of other groups: soluble material is obtained for only approximately half of the proteins. In addition to our interest in structural genomics, we are also interested in the *in vitro* and *in vivo* properties of protein molecules. Two questions then naturally arise: what are the relevant differences in properties between successfully expressed proteins and the rest? Further, how can an understanding of these properties by utilized to greatly improve expression success? We will obtain answers to these questions using a combination of informatics and experimental techniques.

A set of approximately 40 proteins already established as spanning all types of expression outcome – plentiful soluble material, insoluble material, no protein expression in healthy cells, and impaired cell growth, form the basis for the experiments. *E. coli* cellular response to over-expression of these proteins will be investigated using full microarray expression profiles. These data will reveal such factors as specific pathways associated with inclusion body formation, up-regulation of proteases and ribonucleases, differential chaperone expression, and previously unsuspected cellular responses. The primary protein properties influencing expression outcome - stability of the folded state and the rate of folding to that state will be investigated, using microcalorimetry and stopped flow measurements. In the third year of the project, we will test hypotheses generated by these experiments, controlling cell conditions as appropriate, and modifying the properties of the test proteins through mutagenesis.

Results from our own and other structural genomics projects will be stored in a publicly accessible database. These data will be mined for factors that affect expression outcome. We have already discovered relationships between protein family size and expression outcome, and between messenger RNA copy number and expression outcome. Other factors to be investigated include the extent of predicted protein disorder, stability, and folding rate. The results of the data analysis will be used to develop tools for predicting likely expression performance and choosing an optimum expression strategy. In addition to making the data publicly available, we will encourage annotation and discussion of the results, and establish a set of 'challenge proteins' – proteins that have so far not been successfully expressed, but which do not fit the emerging model of protein expression outcome.

The outcome of the project will be a set of informatics and experimental strategies. Informatics will provide a synopsis of all relevant information for a protein, ranking alternative strategies for optimization of production. Possible new strategies include the use of GFP and other reporter fusions to monitor up or down regulation of known and newly discovered cell cellular response proteins; utilization of cellular response to control cell growth; protocols for the design of mutants to improve

expression; inhibition of specific proteins shown to affect outcome; and co-expression of proteins found to enhance outcome.

(Funding for this project is about to begin)

88

High-Throughput Production and Analyses of Purified Proteins

F. William Studier[1,2] (studier@bnl.gov), John C. Sutherland[1,2], Lisa M. Miller[3], and Lin Yang[3]

[1]Biology Department, Brookhaven National Laboratory, Upton, NY; [2]East Carolina University, Greenville, NC; and [3]National Synchrotron Light Source, Brookhaven National Laboratory, Upton, NY

Genome sequences allow access to the proteins of an organism through cloning and expression of the coding sequences. Vectors and protocols designed for high-throughput production of proteins in the T7 expression system in *Escherichia coli* are being developed and will be tested by expressing and purifying proteins of *Ralstonia metallidurans*, a bacterium that tolerates high concentrations of heavy metals and has potential for bioremediation. The vectors are designed to accept PCR products and to donate coding sequences for expression as is, with N- or C-terminal tags, or for co-expression with other coding sequences, as with subunits of protein complexes.

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. Reliable analyses of the state of purified proteins are important for quality assurance in high-throughput production. Stations at the National Synchrotron Light Source analyze proteins by small-angle X-ray scattering (SAXS) to determine size and shape, X-ray absorption microspectrometry to identify bound metals, and Fourier transform infrared (FTIR), UV circular dichroism (CD), linear dichroism (LD) and fluorescence to assess secondary structure and possible intermolecular orientation. An automated sample preparation and loading system to interface between purified proteins in 96-well plates and each of these stations is being constructed to allow high-throughput analyses by these techniques. These assessments of size, shape, secondary structure and metal content of purified proteins will complement analyses such as gel filtration, mass spectrometry and NMR.

Proteomics

89

Ultrasensitive Proteome Analysis of *Deinococcus radiodurans*

Norman J. Dovichi (dovichi@chem.washington.edu)

Department of Chemistry, University of Washington, Seattle, WA

We are developing technology to monitor changes protein expression in single tetrads of *D. radiodurans* following exposure to ionizing radiation. We hypothesize that exposure to ionizing radiation will create a distribution in the amount of genomic damage and that protein expression will reflect the extent of radiation damage.

To test these hypotheses, we have developed the following technologies:

- Fluorescent markers for radiation exposure
- Two-dimensional capillary electrophoresis analysis of the *D. radiodurans* proteome
- Ultrasensitive laser-induced fluorescence detection of proteins separated by capillary electrophoresis

We have generated a number of fully automated two-dimensional capillary electrophoresis separations of proteins extracted from *D. radiodurans*. Figure 1 presents an example, in which the proteins from *D. radiodurans* are first subjected to capillary sieving electrophoretic separation, which is the capillary version of SDS-PAGE using replaceable polymers and which separates proteins based on their molecular weight, with low molecular weight proteins migrating first from the capillary. Fractions are successively transferred to a second capillary, where proteins are separated in a sub-micellar electrophoresis buffer. Components are detected with an ultrasensitive laser-induced fluorescence detector at the exit of that capillary. Over 150 fractions are successively transferred from the first capillary to the second to generate a comprehensive analysis of the protein content of this bacterium. Data are stored in a computer and manipulated to form the pseudo-silver stain image of Figure 1. There are 150 components resolved in this separation.

We have developed a fluorescent DNA damage marker for *D. radiodurans*. We have performed the first successful genetic engineering of this organism to express green fluorescent protein (GFP). We have also engineered the organism to express GFP under control of the recA promoter. This gene is expressed in response to DNA damage, and the GFP fluorescence is produced in response to a variety of DNA damage sources. We have demonstrated the production of GFP under this promoter in *D. radiodurans* in response to ultraviolet radiation and toxin (kanamycin and mitomycin C) exposure, Figure 2. We hope to have data on the gamma radiation response of this system by the time of the conference.

Figure 1.

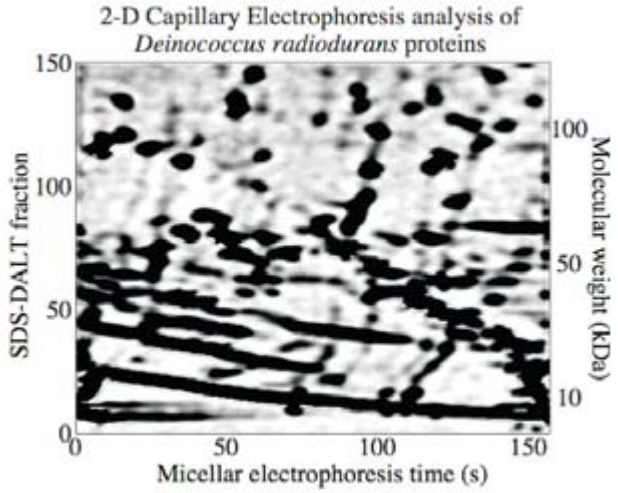
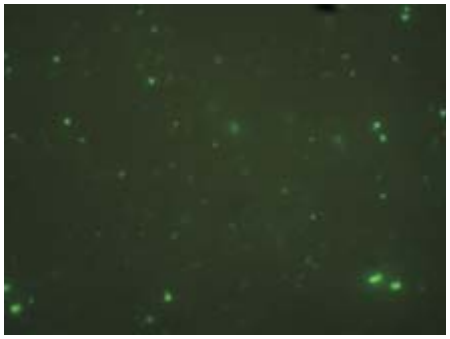


Figure 2. Fluorescent micrographs of recA/GFP engineered D. radiodurans

A. control growth conditions

B. 3-hour exposure to mitomycin C



90

Pilot Proteomics Production Pipeline

Gordon A. Anderson, Mary S. Lipton, Gary R. Kiebel, David A. Clark, Ken J. Auberry, Eric A. Livesay, Vladimir Kery, Brian S. Hooker, Elena S. Mendoza, Ljiljana Paša-Tolić, Matthew Monroe, Margie Romine, Jim Fredrickson, Yuri Gorby, Nikola Tolić, **George S. Michaels** (george.michaels@pnl.gov), and **Richard D. Smith** (dick.smith@pnl.gov)

Biological Sciences Division and Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, WA

Proteomic analysis of biological samples produces large volumes of data from various mass spectrometric technologies. These datasets allow the identification of peptides and proteins as well as allowing quantization of peptide and protein abundances. This research often requires hundreds or thousands of separate MS experiments. These experiments include a liquid chromatographic separation step coupled with both MS and tandem MS experiments. Data analysis tools are then used to perform database searches in order to identify peptides from tandem MS datasets, interpret and extract detected masses and peak elution times from MS datasets, and assign peptide identifications based on those detected masses and times. These complex multistage analyses require tracking of experimental conditions and sample pedigree. Additionally, quality control analysis needs to be performed at several stages during the process to insure instrument performance and sample preparation quality. Applying MS-based proteomics to the determination of the components present in a sample prepared to specifically contain a given bait protein and its specific binding partners requires additional data tracking and automation. This type of research results in large volumes of data as well as many diverse datasets from carefully designed conditions. Our proteomics production pipeline provides an automation platform to monitor, control, acquire, analyze and organize these results. In order to improve the quality of the research results and record critical experiment and analysis metadata, PNNL has developed an automated proteomics pipeline that includes the following key components:

- **Automated Data Management and Analysis**.

  Sample, analysis, and experimental process data are recorded throughout the pipeline by means of 3 key features. The first is a prototype LIMS system that will cover the experiment design process, gathering process data as well as QA/QC data, and then track those samples until they are ready for MS analysis. Next is commercial freezer monitoring software that will track the movements of specific samples into and out of the freezers, allowing for better inventory control and sample tracking. Barcodes will be used throughout the pipeline to uniquely and quickly identify a sample. Once the samples are ready for MS analysis, they will be handed off to the PRISM system and tracked from there. Analysis of raw mass spectrometer data includes several processing steps involving a combination of commercial data analysis tools and applications developed at PNNL. Automation of the analysis pipeline is performed using our Proteomics Research Information Storage and Management system (PRISM). PRISM automates the capture of data from the mass spectrometers, data reduction of raw data to tables of detected peptides from MS/MS datasets, and tables of detected masses from MS datasets. PRISM then further analyzes this reduced data to develop database tables containing identified peptides and proteins to be used by higher order analysis steps. PRISM allows the users to

monitor the status of analysis and to schedule and track samples through this portion of the proteomics pipeline. These three systems will be connected together through the common use of sample identifiers represented by barcodes.

- **Automation**

  High throughput requires automation; additionally automation provides better control of the process and improves repeatability. Many of the labor intensive and critical aspects of the process are being automated. These automation systems include, protein complex sample preparation, peptide digestion sample preparation, LC-MS and LC-MSMS experiment control. These automation steps include integration with the LIMS to define processing parameters and track the samples as they progress through the system.

- **Data Abstraction Layer (DAL)**

  The DAL is middleware that will provide a level of abstraction for any data storage system in the proteomics pipeline (LIMS, Freezer Software, PRISM, etc). It will provide a generic interface for building tools and applications that require access to the experimental data and analysis results. It will also allow the pipeline data to be extended without making changes in the manner in which an application already looks at the data. For example, it could be used to facilitate a query performed utilizing proteomic data originating from both PNNL and ORNL. The DAL will be used to provide an interface to the pipeline data as required by selected bioinformatics/analysis tools.

Development of the production pipeline lays the foundation for high throughput proteome analysis. This system tracks samples, metadata and raw data for all steps of the process and provides this data to bioinformatics tools through a standard interface. This allows evolution of the PRISM and LIMS system while insulating bioinformatics tools from these changes through the DAL.

91

Characterization of Microbial Systems by High Resolution Proteomic Measurements

Mary S. Lipton (mary.lipton@pnl.gov), Ljiljana Paša-Tolić, Matthew E. Monroe, Kim K. Hixson, Dwayne A. Elias, Margie F. Romine, Yuri A. Gorby, Ruihua Fang, Heather M. Mottaz, Carrie D. Goddard, Nikola Tolić, Gordon A. Anderson, Richard D. Smith, and Jim K. Fredrickson

Biological Sciences Division and Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, WA

Collaborators: Michael Daly (Uniformed Services University of the Health Sciences), Timothy Donohue (University of Wisconsin), Samuel Kaplan (UT-Houston Medical School), and Derek Lovley (University of Massachusetts)

Developing a systems-level understanding of how cells function requires technologies that are capable of making global measurements of protein abundances (i.e., the "proteome"). At PNNL, new technologies based primarily upon high resolution separations combined with Fourier transform ion cyclotron resonance mass spectrometry have been developed and applied to obtain quantitative and high throughput global proteomic measurements of microorganisms of interest to DOE mission

areas. Among the microorganisms of interest are *Shewanella oneidensis* MR1, *Deinococcus radiodurans* R1, *Rhodobacter sphaeroides* and *Geobacter sulfurreducens*. Significant progress has been made addressing biological questions associated with each of these organisms using high resolution proteomic measurements of cells, and fractions thereof, cultivated under varying conditions.

S. ondeidensis MR-1, a Gram-negative, facultative anaerobe and respiratory generalist, is of interest to DOE because it can oxidize organic matter using metals such as Fe(III) or Mn(III,IV) as the electron acceptors. It can also reduce soluble U(VI) to the insoluble U(IV) form. This ability to reduce U prevents further U mobility in groundwater and subsequent contamination of down-gradient water resources. Microbial reduction shows significant promise for the *in situ* bioremediation of subsurface environments contaminated with U, Tc, and toxic metals such as chromate. A recent revised annotation of the *S. oneidensis* genome suggested a number of changes in the proteins predicted to be expressed by the organism. Using the extensive mass tag database we assembled for this organism and highly stringent criteria for peptide/protein identification, we have for example, analyzed proteome data generated from 172 tryptic digests of *S. oneidensis* MR-1 cellular proteins for the occurrence of peptides associated with proteins less than 101 amino acids in length or that were added to the genome annotation after its initial deposit in Genbank. The mass tag approach also has enabled qualitative experiments to determine the presence or absence of particular proteins in samples as well as quantitative experiments to determine the changes in protein expression upon changes in culture condition. Strategies that use both stable isotope labeling and MS peak intensities of these mass tags provide the basis for quantitation and have been applied to collaborative experiments designed to determine changes in protein expression in cells grown under aerobic and anaerobic conditions.

Similar to *S. oneidensis, Geobacter sulfurreducens* is a dissimilatory metal-reducing bacterium that can reduce soluble U(VI) to insoluble U(IV). Other projects under the DOE Microbial Genome Program have already sequenced the *G. sulfurreducens* genome and have initiated a functional genomics study to elucidate genes of unknown function in this organism. Proteomic efforts with this microorganism are currently focused on creating a mass tag database. Initial global protein expression determinations have shown protein expression in most functional categories as assigned by TIGR. Early uses of the database have centered on determining proteins contained within the membrane of the organism; future studies will be extended to include *Geobacter* dominated microbial communities.

The most significant characteristic of *D. radiodurans* is its ability to resist the lethal effects of DNA damaging agents such as ionizing radiation, UV radiation, hydrogen peroxide and desiccation. The capacity for survival after severe DNA damage at such high levels of ionizing radiation is currently unclear and may be the result of unusually efficient repair and/or protection mechanisms. We utilized the extensive mass tag database developed for *D. radiodurans* and applied a combination of stable isotope labeling and MS peak intensities to determine quantitative changes in protein expression for the organism (1) grown in rich and minimal media, (2) exposed to an acute dose of radiation, and (3) cultured in the presence of chronic radiation.

Rhodobacter sphaeroides 2.4.1 is a-3 purple nonsulfur eubacterium with an extensive metabolic repertoire. Under anaerobic conditions, it is able to grow by photosynthesis, respiration and fermentation. By quantitative measurement of the proteome of *R. sphaerodies* cultured under specific growth conditions, we aim to identify the proteins involved in the different metabolic pathways. For the initial mass tag database, the organism was cultured under both aerobic and photosynthetic conditions,

and differences in the proteins expressed under the two conditions are being determined. Additionally, cellular fractions of these organisms cultured under both aerobic and photosynthetic cell states have been. Photosynthetic cells have been fractionated into 5 relatively discreet fractions (cytosol, periplasm, inner membrane, photosynthetic membrane and outer membrane) and the aerobic cells have been fractionated into 4 relatively discreet fractions (cytosol, periplasm, inner membrane, and outer membrane) in an effort to determine protein localization in the cell. We will be able to determine the changes in localization of specific proteins upon change in cellular state.

The accuracy and precision in which to make proteomic measurements as described above is intricately linked with the instrumentation in which the measurements are made as well as the efficiency of the sample processing methods. Advances in automation of sample processing will reduce variation between digested samples. Additionally, improved methods for quantitation and the application of increasingly sophisticated bioinformatics tools for data analysis will enormously improve the types and quality of the proteomic data available in the future.

92

Advanced Technologies and Their Applications for Comprehensive and Quantitative Microbial Proteomics

Richard D. Smith (dick.smith@pnl.gov), Mary S. Lipton, Ljiljana Paša-Tolić, Gordon A. Anderson, Yufeng Shen, Matthew Monroe, Christophe Masselon, Eric Livesay, Ethan Johnson, Keqi Tang, Harold R. Udseth, and David Camp

Biological Sciences Division and Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, WA

Essential to realizing the ambitious goals of the Genomics:GTL (GTL) Program is the ability to characterize the broad array of proteins potentially expressed by both individual microbes and complex microbial communities. While recent technological advances are laying the foundation for proteomics approaches that provide more effective, more comprehensive and higher throughput protein measurements, the challenges associated with making truly useful comprehensive proteomics measurements are considerable. Among the challenges are the abilities to identify and quantitate large sets of proteins from highly complex mixtures with components of interest having relative abundances potentially spanning more than six orders of magnitude, that vary broadly in chemical and physical properties, that can have transient and low levels of modifications, and that are subject to endogenous proteolytic processing. Additionally, the utility of proteomics data depends significantly on the quality of the data – both the confidence of protein identifications as well as the quantitative usefulness of the data.

The proteomics technology and approaches developed at PNNL under DOE support employ high resolution nano-scale, ultra-high pressure capillary liquid chromatography (cLC) separations combined with extremely high accuracy mass measurements obtained with Fourier transform ion cyclotron resonance (FTICR) mass spectrometry. The quality of these measurements allow one to identify and designate accurate mass and (cLC) time (AMT) peptide tags that are markers for proteins. These AMT peptide tags can be used in subsequent mass spectrometric measurements, avoiding the throughput limitations associated with routine peptide

identification using tandem mass spectrometry. This approach enables fundamentally greater throughput and sensitivity for proteome measurements. Currently, our prototype FTICR proteomics production line is running in high throughput mode "24/7". A new capability for "data-dependent" tandem mass spectrometry allows otherwise uncharacterized peptides to be selected and characterized directly in the FTICR (i.e., peptides that have not been previously identified and designated as AMT tags). When coupled with stable isotope labeling to allow the direct analysis of two differently labeled samples, this technology identifies those peptides that change significantly in abundance. Additional new developments have significantly extended the dynamic range of measurements to approximately six orders of magnitude and are now providing the capability for proteomic studies from very small cell populations, and even to the level approaching that of single cells.

Under DOE support, microbial systems we have extensively characterized include *Deinococcus radiodurans, Shewanella oneidensis, Rhodobacter sphaerodies* and *Geobacter sulfurreducens.* We have developed extensive AMT peptide tag databases for the first two microbes and are in process of developing databases for the latter two. In addition, significant efforts have been made towards characterizing the proteomes of *Rhodopseudomonas palustris*, *Synechocystis*, *Borrelia burgdorferi*, *Desulfovibrio vulgaris*, and *Methanosarcina barkeri*. We have successfully incorporated the use of protein and peptide fractionation in the initial mass tag identification step (based on conventional tandem mass spectrometry in an ion trap), which has increased the dynamic range of these experiments and thus the number of AMT peptide tags. This type of proteomic data can be used in a variety of experiments, ranging from quantitative studies comparing one culture condition to another, to protein localization experiments where cellular fractions are analyzed for their protein content. The use of either stable isotope labeling or MS peak intensities of these AMT peptide tags provides the basis for quantitation. The use of peak intensities potentially circumvents the need for expensive stable isotope labeling methods, and provides a basis for obtaining quantitative information for non-culturable organisms and microbial communities.

A significant challenge for proteomics studies is the immense quantity of data that must be managed and effectively processed and analyzed in order to be useful. Thus, a key component of our program involves development of the informatics tools necessary to make the data more broadly available to the research community and to extract knowledge and new biological insights from complex data sets. A new software tool called VIPER has been developed to automatically process FTICR data sets, which has streamlined data processing. VIPER works with the overall PRISM data management system developed at PNNL to automatically extract and coordinate the use of various types of pertinent information in the application of AMT tags, in addition to managing routine functions, such as FTICR data archiving.

This presentation will describe development and application of new technologies for global proteome measurements that are orders of magnitude more sensitive and faster than previous technologies and that can address many of the needs of the GTL program. The status of the technology will be described in the context of applications, and the basis for extending the applications to more complex microbial communities will also be described.

93

New Developments in Peptide Identification from Tandem Mass Spectrometry Data

William R. Cannon[1] (William.cannon@pnl.gov), Kristin H. Jarman[2], Alejandro Heredia-Langner[2], Douglas J. Baxter[3], Joel Malard[2], Kenneth J. Auberry[4], and Gordon A. Anderson[4]

[1]Statistics and Quantitative Sciences, Computational Sciences and Mathematics Division; [2]Molecular Sciences Computing Facility, Environmental Molecular Sciences Laboratory; [3]Computational Biosciences Group, Biology Division; and [4]Instrument Development Lab, Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, WA

We present a flexible statistical framework for identification of peptides from the tandem mass spectrometry data. The statistical model is based on a two-sided hypothesis test that compares the likelihood that a spectrum is due to a specific peptide to the likelihood that the peptide arose by chance. The likelihoods are computed from the probability of occurrence of peptide fragments from the parent peptide. These probabilities are empirically derived from fragmentation patterns from a training set of 16,134 spectra of varying charge, composition and length. As a result, a fragmentation model is developed from which model spectra are generated for comparison to real spectra and scoring peptides. The code for the analysis runs on both serial and parallel computers. The statistical model is evaluated on an independent data set of 19,000 spectra using the parallel version of the code on a large Linux cluster.

In addition, we present a sequence optimization approach as an alternative to de novo peptide analysis to reconstruct amino acid sequences of peptides. The sequence optimization can potentially overcome some of the most problematic aspects associated with *de novo* analysis of real MS/MS data such as incomplete or unclearly defined peaks and may prove to be a valuable tool in the proteomics field. We assess the performance of our algorithm under conditions of perfect spectral information, in situations where key spectral features are missing, and using real MS/MS spectral data. The prototype algorithm we use performs well under these situations.

Metabolomics

94

New, Highly Specific Vibrational Probes for Monitoring Metabolic Activity in Microbes and Microbial Communities

Thomas Huser (huser1@llnl.gov), Chad Talley, Allen Christian, Chris Hollars, Ted Laurence, and Steve Lane

Lawrence Livermore National Laboratory, Livermore, CA

We are currently developing a set of new, stable and highly specific intra- and extracellular probes that can monitor metabolic activity inside and in the immediate environment of individual prokaryotic cells. Our sensing technology makes use of vibrational probes (functionalized gold/silver nanoparticles) that monitor the chemical levels inside single microbes with nanometer resolution. These probes consist of specific marker molecules for metabolic byproducts that are chemically linked to the surface of metal nanoparticles with diameters ranging from 40-100 nm. The response of these marker molecules to changes in their local environment can be probed through changes in their characteristic Raman spectrum inside single microbes and in microbial communities. These probes are made of biocompatible and inert materials, they are easy to probe by highly sensitive micro-Raman spectroscopy, and they are very bright and photostable and provide quantitative information about the concentration of metabolic byproducts in their immediate environment. We plan to develop these probes for a range of metabolites and demonstrate their applications in cultured and uncultured microbial communities.

We also demonstrate the use of laser-tweezers Raman spectroscopy, where individual cells are optically suspended in a highly focused laser beam, which at the same time characterizes the chemical activity of the cells by their Raman spectrum. We demonstrate how this capability can be used to distinguish between different cells or monitor their chemical response to external stimuli.

95

New Technologies for Metabolomics

Jay D. Keasling (jdkeasling@lbl.gov), Carolyn Bertozzi, Julie Leary, Michael Marletta, and David Wemmer

Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA

Microorganisms have evolved complex metabolic pathways that enable them to mobilize nutrients from their local environment and detoxify those substances that are detrimental to their survival. Metals and actinides, both of which are toxic to microorganisms and are frequent contaminants at a number of DOE sites, can be immobilized and therefore detoxified by precipitation with cellular metabolites or by reduction using cellular respiration, both of which are highly dependent on cellular metabolism. Improvements in metal/actinide precipitation or reduction require a

thorough understanding of cellular metabolism to identify limitations in metabolic pathways. Since the locations of bottlenecks in metabolism may not be intuitively evident, it is important to have as complete a survey of cellular metabolism as possible. Unlike recent developments in transcript and protein profiling, there are no methods widely available to survey large numbers of cellular metabolites and their turnover rates simultaneously. The system-wide analysis of an organism's metabolite profile, also known as "metabolomics", is therefore an important goal for understanding how organisms respond to environmental stress and evolve to survive in new situations, in determining the fate of metals and actinides in the environment, and in engineering or stimulating microorganisms to immobilize these contaminants.

The goals of this project are to develop methods for profiling metabolites and metabolic fluxes in microorganisms and to develop strategies for perturbing metabolite levels and fluxes in order to study the influence of changes in metabolism on cellular function. We will focus our efforts on two microorganisms of interest to DOE, *Shewanella oneidensis* and *Geobacter metallireducens*, and the effect of various electron acceptors on growth and metabolism. Specifically, we will (1) develop new methods and use established methods to identify as many intracellular metabolites as possible and measure their levels in the presence of various electron acceptors; (2) develop new methods and use established methods to quantify fluxes through key metabolic pathways in the presence of various electron acceptors and in response to changes in electron acceptors; (3) perturb central metabolism by deleting key genes involved in respiration and control of metabolism or by the addition of polyamides to specifically inhibit expression of metabolic genes and then measure the effect on metabolite levels and fluxes using the methods developed above; and (4) integrate the metabolite and metabolic flux data with information from the annotated genome in order to better predict the effects environmental changes on metal and actinide reduction.

Recently, microorganisms have been explored for metal and actinide precipitation by secretion of cellular metabolites that will form strong complexes or by reduction of the metal/actinide. A complete survey of metabolism in organisms responsible for metal and actinide remediation, parallel to efforts currently underway to characterize the transcript and protein profiles in these microorganisms, would allow one to identify rate limiting steps and overcome bottlenecks that limit the rate of precipitation/reduction.

Not only will these methods be useful for bioremediation, they will also be useful for improving the conversion of plentiful renewable resources to fossil fuel replacements, a key DOE mission. For example, the conversion of cellulosic material to ethanol is limited by inefficient use of carbohydrates by the ethanol producer. Identification of limitations in cellulose metabolism and in products other than ethanol that are produced during carbohydrate oxidation could lead to more efficient organisms or routes for ethanol production – metabolomics is the key profile to identify these rate-limiting steps.