Section 3

Protein Production and Characterization

21

Genes to Proteins: Elucidating the Bottlenecks in Protein Production

M. Hadi[1]* (MZhadi@sandia.gov), K. Sale[1], J. Kaiser[1], J. Dibble[1], D. Pelletier[2], J. Zhou[2], B. Segelke[3], M. Coleman[3], and L. Napolitano[1]

[1]Sandia National Laboratories, Livermore, CA; [2]Oak Ridge National Laboratory, Oak Ridge, TN; and [3]Lawrence Livermore National Laboratory, Livermore, CA

The ability to produce proteins is currently a major biological, physical, and computational challenge in protein research. Given a standard set of conditions, less than 30% of any given genome is expressible in a recombinant host. Protein expression requires complex, lengthy procedures, and specific proteins commonly require individual strategies for optimal expression. Standard bench-level procedures for protein production (expression and purification) do not exist. This lack of validated processes leads to a lengthy search for correct vector, host, expression and purification conditions to yield protein in milligram amounts.

A recent paradigm shift in life science research is from characterizing single genes or proteins (over an investigator's career) to studying whole genomes, proteomes or specific pathways in single "experiments" in a few months. It has been known for some time that while it is quite straightforward to clone and over-express a protein of interest to visualize it on an SDS-PAGE gel. However, it is a completely different matter to obtain purified protein in amounts sufficient for structural and functional studies.

At SNL we have developed successful recombination-based and directed cloning methods for generating large numbers of expression constructs for protein expression and purification. This medium-throughput pipeline is now being automated. This pipeline was initially setup to process targets that included the stress response genes from *D. vulgaris* (GTL funded project). We are using this pipeline to express several hundred ORFs from microbes of DOE relevance (*D. vulgaris, S. oneidensis and R. paulustris*). Each ORF is being expressed using in vivo (*E. coli*) and three in vitro systems (two different prokaryotic and one eukaryotic system). Using the data generated from our expression runs, as well as data generated by other researchers, we are developing computational tools to predict optimal expression system and conditions based on primary sequence and other known properties of the protein. Initially, the prediction software will be specific for prokaryotic systems. However, it will be designed to easily generalize to mammalian systems once training data for mammalian systems is acquired.

* Presenting author

22

High-Throughput Production and Analyses of Purified Proteins

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

[1]Brookhaven National Laboratory, Upton, NY and [2]East Carolina University, Greenville, NC

This work is aimed at improving the efficiency of high-throughput protein production from cloned coding sequences and developing a capacity for high-throughput biophysical characterization of the proteins obtained. Proteins of *Ralstonia metallidurans*, a bacterium that tolerates high concentrations of heavy metals and has potential for bioremediation, are being produced to test and improve the efficiency of protein production in the T7 expression system in *Escherichia coli*. Auto-induction greatly simplifies protein production, as cultures can simply be inoculated and grown to saturation without the need to monitor culture growth and add inducer at the proper time. New vectors allow maintenance and expression of coding sequence for proteins that are highly toxic to the host. Vectors having a range of expression levels have also been made and will be tested for whether tuning the expression level can improve the production of soluble, well folded proteins.

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

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

* Presenting author

23

A Combined Informatics and Experimental Strategy for Improving Protein Expression

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

Center for Advanced Research in Biotechnology, Rockville, MD

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

Informatics methods have now been used to examine a wide range of factors potentially affecting soluble expression, including protein family size, native expression level, low complexity sequence, open reading frame validity, amyloid propensity and inherent disorder. Of these, the most significant ones affecting expression outcome are native expression level, family size, and inherent disorder. The relevance of disorder is being investigated further.

In the first year of the project, we established that the expression of a set of host proteins are consistently upregulated under insoluble protein production conditions, Building on that finding, we are developing reporter constructs utilizing the genes for green fluorescent protein and luciferase, both of which can be assayed by spectrophotometer in intact cells. These reporters will be used to screen a large variety of growth conditions for improved expression of various insoluble recombinant proteins.

Protein stability measurements on a set of proteins using differential scanning calorimetry have been extended, using chemical denaturation with guanidine hydrochoride. So far, the new data support the earlier suggestion that stability is not major factor in determining soluble expression.

This project is supported by Genomes to Life award DE-FG02-04ER63787.

24

High-Throughput Optimization of Heterologous Proteome Expression

Robert Balint (rfbalint@sbcglobal.net) and Xiaoli Chen

CytoDesign, Inc., Mountain View, CA

High-throughput genomic sequencing has begun to reveal the universe of microbial metabolic capabilities, which can be harnessed to provide new renewable sources of energy, bioremediation, and control of carbon cycling and sequestration, as well as new tools for pharmaceutical discovery, biomolecular synthesis and production, and industrial processes, among others. However, realizing these potentialities will require exhaustive structural and functional characterization of thousands of proteome constituents. Conventional methods for these tasks are so time- and labor-intensive that many years may be required for comprehensive characterization of proteomes of interest, and the costs may be prohibitive. Thus, new methods are urgently needed to accelerate these processes. For structure/function studies *in vivo* and *in vitro*, proteome constituents must be expressed in one or more suitable hosts, accumulating to functional levels in soluble form in their native conformations. They must be purified in sufficient quantities for structural determinations, and also for isolation of analytical reagents such as antibodies. However, many proteome members fail to meet these requirements for one or more of four main reasons: (1) failure to fold before aggregation, (2) failure to fold before proteolytic turnover, (3) instability in the host of choice, or (4) toxicity to the host of choice. An even larger percentage fail to express well enough to provide sufficient material for one or more essential analytical procedures.

The present project addresses the proteome expression problem by adapting selectable reporter systems in *E. coli* for use in high-throughput optimization of heterologous protein expression using peptide chaperones, *i.e.*, small peptides genetically linked to the amino or carboxyl termini of proteome members, which are selected from small libraries for their ability to chaperone folding into native conformations to permit high levels of soluble expression. The chaperone selection systems are based on quantitative coupling of soluble expression to the activity of a reporter which confers a selectable phenotype on the cells. Systems are available for optimization of both secretory and cytoplasmic expression, and the systems permit the optimization of multiple proteins (tens to hundreds) simultaneously.

For convenience and efficiency, proteomes of interest are being optimized for high-level secretory expression in the *E. coli* periplasm. High-level periplasmic expression allows simple recovery and rapid, efficient one-step affinity purification from growth media and/or osmotic shock extracts of continuous or fed-batch cultures, which can be readily scaled to >100 mg/L yields. Periplasmic expression also allows for direct high-throughput selection of antibody affinity reagents for proteome proteins without the need for prior protein purification. The selection system of choice for high-throughput optimization of periplasmic expression is illustrated in the figure below. In this system the reporter is an unstable circular permutation of β-lactamase (β-lacCP), in which a flexible polypeptide linker has been inserted between the native termini, and new "break-point" termini have been introduced at a site within a loop on the surface of the enzyme. The β-lacCP can be fully activated by adding cysteine residues (SH) to the break-point termini, which allow formation of a disulfide bond (SS) to stabilize the locus around the break-point. However, to couple β-lacCP activation to proteome protein expression, a leucine zipper docking mechanism was introduced, in which one cysteine on the CP is replaced by one helix of the zipper, and the other zipper helix, bearing the second cysteine, is linked to a proteome protein of interest (Proteomer). When the

* Presenting author

Proteomer and the β-lacCP are co-expressed in the periplasm of the same cells, formation of the leucine zipper allows formation of the β-lacCP-activating disulfide in proportion to the expression level of the Proteomer. When the Proteomer is equipped with a random sequence library of typically 3-9 residues at its N-or C-terminus, some peptides will (1) protect the Proteomer from aggregation, (2) accelerate folding, (3) facilitate translocation, and/or (4) stabilize the Proteomer, so that the cells expressing these peptides will grow and can be isolated on otherwise non-permissive antibiotic concentrations.

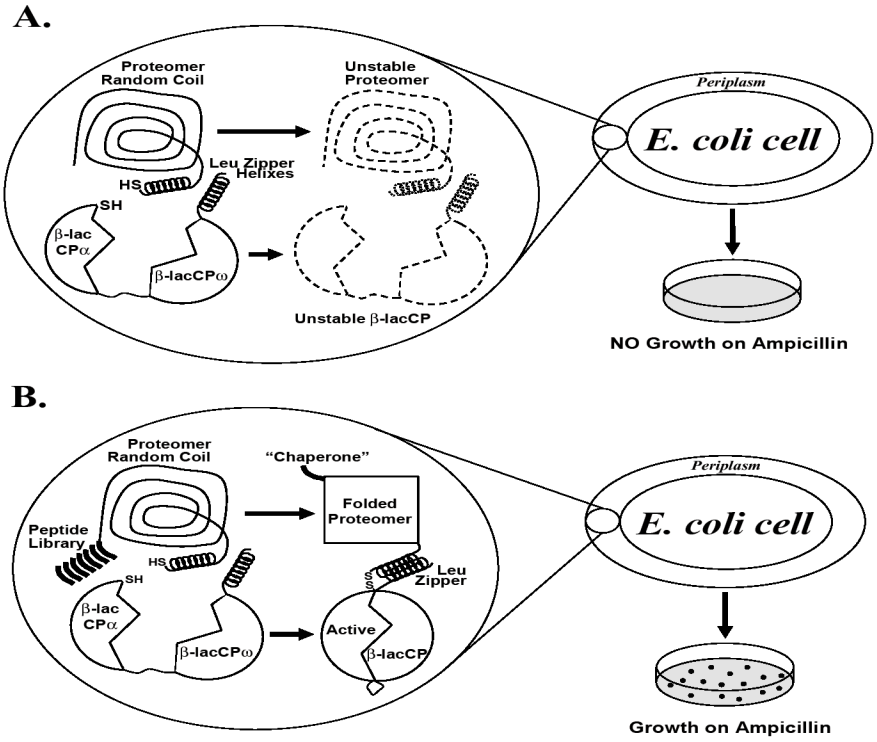

Figure 1. The system is being used to optimize the expression of the proteome of *Pseudomonas putida*, one of nature's most versatile microbes. *P. putida* has the most genes of any known species for breaking down aromatic hydrocarbons, like TNT, and it also reduces a broad spectrum of toxic metals. It encodes at least 80 oxidative reductases for biomass decomposition, and has hundreds of genes for sensing chemicals in the environment. So far, 13 open reading frames (ORFs) selected at random have been expressed in the system with and without N-terminal hexapeptide libraries. Non-permissive antibiotic concentrations were determined for each ORF, and the ORFs were pooled into 3 groups for peptide chaperone selection. At least one peptide was obtained for each ORF, which substantially increased expression, in some cases to >100 mg/L. An added feature of the system is that a suppressible stop codon can be inserted between ORF and leucine zipper helix, so that selections can be performed in a suppressor host, and then selected ORF-peptide constructs can be retransformed into a non-suppressing host for expression without the leucine zipper helix without the need for subcloning.

25

Development of Genome-Scale Expression Methods

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

Argonne National Laboratory, Argonne, IL

We are developing novel cellular and cell-free technologies to optimize the expression of cytoplasmic, periplasmic/secreted proteins and protein domains targets from prokaryotic and eukaryotic organism of programmatic interest. The program incorporates technology development and production components and focuses on the use of automated systems and implementation of robotic methods (96-well plate based) to expand the boundaries of current high throughput technology. The technology development aspect applies various expression strategies to target groups and documents the success rate for production of clones validated for expression of a soluble protein target. One example of this component is a project to document expression/solubility outcomes for targets directed to the cytoplasm or periplasmic compartment of *Escherichia coli*. The primary target groups being evaluated in the initial round include the set of periplasmic proteins from *Shewanella oneidensis* as well as a set of simple architecture membrane proteins from *Geobacter sulfurreducens*. The simple architecture membrane proteins contain a membrane anchor or a predicted single membrane spanning helix. For expression/solubility screening of these targets, individual domains adjacent to the membrane region are cloned and targeted to the cytoplasmic or periplasmic compartments.

An important aspect of the production component is the transfer of data and physical resources to experimental labs. The development of target lists and the preferred resources for production of the targets are linked to GTL collaborators with their input and participation solicited at multiple levels including project design, workflow management, and resource distribution. This process involves the generation of specialized vectors to meet experimental requirements [e.g. tag vectors enabling pulldowns for interaction screening or activity screening, or in vivo interaction screening]. One outcome of these efforts is the generation of an expression clone resource for protein targets from various organisms. Our current *E. coli* expression clone library contains over a thousand clones that have been screened for expression and solubility with more than 600 clones that have been validated for successful expression of a soluble protein. The current list of expression clone resource available for distribution can be found on the project website (http://www.bio.anl.gov/combinatorialbiology/GTL.htm). We have extended the scope of available resources based on the need for purified and characterized proteins. In a pilot project to survey requirements for systematic production of proteins, we have purified and distributed >400 purified proteins in mg quantities. Our experience suggests these represent a valuable resource but that considerable effort needs to be expended to define the requirements for large scale protein production.

* Presenting author

26

A Novel Membrane Protein Expression System for Very Large Scale and Economical Production of Membrane Proteins

Hiep-Hoa T. Nguyen[1]* (hiephoa@its.caltech.edu), Sanjay Jayachandran[1], Randall M. Story[1], Sergei Stolyar[3], and Sunney I. Chan[2]

[1]TransMembrane Biosciences, Pasadena, CA; [2]California Institute of Technology, Pasadena, CA; and [3]University of Washington, Seattle, WA

A majority of membrane proteins are very difficult to obtain in any significant quantities, even at milligrams scale since their natural biosynthesis levels often are very low and currently available protein expression systems are not effective for membrane proteins. With the completion of several genome sequencing projects, many large-scale efforts are under way to understand the protein products including the DOE Genome-to-Life project. The lack of effective method for preparative-scale membrane protein synthesis will hamper progress toward a complete understanding of the proteome and prevent us to take full advantage of available sequences, especially of membrane proteins with medicinal importance.

Although currently available *in vivo* protein expression systems are very powerful, capable of producing gram quantities of soluble proteins, their applications to membrane protein synthesis have yielded very poor results. We are working to develop a powerful yet economical host/vector membrane protein expression system utilizing a group of bacteria capable of synthesizing very large quantities of membrane proteins. A series of expression vectors has been created and can be used to express a variety of membrane proteins in these bacteria. Concurrently, other molecular biology tools/protocols are also being developed to genetically engineer these organisms through extensive genome modifications in order to enhance the yield of correctly folded and functional recombinant membrane proteins. Preliminary data from these experiments will be presented.

27

High-Throughput Methods for Production of Cytochromes *c* from *Shewanella oneidensis*

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

Argonne National Laboratory, Argonne, IL

One of major challenges of post-genomic biology is the development of high-throughput methods for heterologous production of proteins from newly sequenced genomes. The capability for comprehensive production of proteins is an essential component of the systems biology focus of the Genomics:GTL program. Cytochromes *c*, where heme is covalently bound to the polypeptide chain, represent a challenge for heterologous expression systems, since the nascent apoprotein must undergo correct post-translational modification (heme attachment). Our work addresses two major challenges presented by cytochromes *c* with respect to high-throughput production – plate-based detection techniques and the development of a vector resource suitable for periplasmic targeting and efficient expression of cytochromes *c*.

Polyhistidine tags are routinely used for detection of expression and solubility levels in a high-throughput screening methods for protein expression and solubility. However, these tags can interfere with heme attachment and/or folding, especially for multiheme cytochromes, and, therefore, alternative methods to screen for expression and solubility of cytochromes would be useful. We developed a plate-based assay to screen multiple clones for their ability to express *c*-type cytochromes. The assay takes advantage of intrinsic peroxidase activity of heme that can be monitored using commercially available substrates for ELISA. Even though, in the majority of *c*-type cytochromes heme is coordinated by two proteinaceous ligands and not accessible for a molecule of hydrogen peroxide under native conditions, peroxidase activity is preserved upon denaturation of the protein in 6 M guanidine chloride. The method is sensitive enough to detect cytochrome concentrations < 10 μg/ml, which is at least an order of magnitude lower than the concentration of an average recombinant protein in a cell lysate.

The other critical requirement is the availability of high throughput compatible vectors that target nascent polypeptides to the periplasm since targeting to this compartment is necessary for heme attachment to apocytochromes. We designed two families of high throughput compatible vectors that incorporate ligation independent cloning (LIC) sites. One family features a novel LIC-site that allows cloning of targets without the addition of extra residues to the N-terminus (thus preserving a wild-type N-terminus upon the leader peptide cleavage). We have also generated two variants of this vector which enable constitutive co-expression of different periplasmic chaperones. The other family of LIC compatible vectors allows cloning of target genes as fusions to different carrier proteins to facilitate folding and purification. We have used this vector resource to express 30 genes from *Shewanella oneidensis* coding for cytochromes *c* or cytochromes *c*-type domains predicted to have 1-4 hemes. After DNA sequence confirmation, expression and solubility levels were evaluated by SDS-PAGE and the transferred gel pattern stained for heme proteins. Large scale purification and preliminary physico-chemical characterization of individual cytochromes are currently being undertaken.

* Presenting author

28

Towards the Total Chemical Synthesis of Helical Integral Membrane Proteins

Erik C.B. Johnson* (erikj@uchicago.edu) and **Stephen B.H. Kent**

University of Chicago, Chicago, IL

Polytopic helical integral membrane proteins represent an important class of proteins in the cell, yet our understanding of how they function on a molecular level remains elementary due to the inherent difficulties in producing and handling them in their functional forms. ***Chemical protein synthesis*** (CPS) potentially offers an alternative route to the production of integral membrane proteins in quantities sufficient for biophysical studies, yet has been hampered by the difficulty in synthesizing, handling, and purifying peptides that contain transmembrane (TM) domains. These peptides are largely ***insoluble*** in aqueous and mixed organic/aqueous solvents, and show a strong tendency to aggregate. To address this issue, we are developing *reversible backbone protection* as a means to chemically render TM peptides soluble and enable the use of synthetic TM peptides, in conjunction with native chemical ligation, for the total chemical synthesis of helical integral membrane proteins.

29

Selecting Binders Against Specific Post-Translational Modifications: The Sulfotyrosine Example

John Kehoe[3], Jytte Rasmussen[2], Monica Walbolt[2], Carolyn Bertozzi[2], and **Andrew Bradbury**[1]* (amb@lanl.gov)

[1]Los Alamos National Laboratory, Los Alamos, NM; [2]University of California, Berkeley, CA; and [3]Centocor, Horsham, PA

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

We designed a number of different selection strategies, using peptides containing the sulfotyrosine modification as positive selectors in the presence of an excess of the non-modified peptide as blocking agent. We screened almost eight thousand clones after two or three rounds of selection and identified a single scFv able to recognize tyrosine sulfate in multiple sequence contexts. Further analysis shows that this scFv is also able to recognize naturally sulfated proteins in a sulfation dependent fashion.

30

Progress on Fluorobodies: Intrinsically Fluorescent Binders Based on GFP

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

Los Alamos National Laboratory, Los Alamos, NM

Antibodies are the most widely used binding ligands in research. However, they suffer from a number of problems, especially when used in molecular diversity techniques. These include low expression levels, instability and poor cytoplasmic expression, as well the inability to detect binding without the use of secondary reagents. The use of GFP as a scaffold would resolve many of these problems. However, due to the destabilization of GFP folding upon the insertion of extraneous sequences, it has not been possible to use standard GFP as an effective scaffold. Initial attempts to insert diversity into an extremely stable form of GFP (Superfolder GFP) and use phage display were unsuccessful. We have now overcome these problems and have succeeded in selecting GFP based binders which preserve both fluorescence and binding activity. These bind their targets specifically as shown by ELISA, FLISA and flow cytometry, with affinities (measured using surface plasmon resonance) in the nanomolar range.

Fluorescent proteins only become fluorescent when correctly folded. This property becomes extremely useful in the design, selection, screening and use of fluorescent binders, in particular:

- Making libraries; diversity compatible with folding can be selected, screened or monitored
- Monitoring the selection process
- Analyzing expression and affinity of selected fluorescent binders
- Assessing functionality: if it is fluorescent, it is functional
- As a downstream detection signal in e.g. immunofluorescence, FLISA, flow cytometry, biosensors

These binders hold tremendous potential in many different fields, including proteomics and high throughput selection projects, such as the GTL protein production and affinity reagents facility.

* Presenting author

31

High Throughput Screening of Binding Ligands using Flow Cytometry

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

Los Alamos National Laboratory, Los Alamos, NM

Phage display libraries represent a relatively easy way to generate binding ligands against a vast number of different targets. Although in principle, phage display selection should be amenable to automation, this has not yet been described and present selection protocols are far from high throughput. We have examined the selection process in a systematic approach and automated most of the individual steps. Selection is carried out in the microtiter format using 24 targets as the individual selection lot size. We are transitioning from screening by ELISA to using a flow cytometric approach, in which the reactivity of individual antibody clones for numerous different target parameters can be examined simultaneously. This is done by labeling binders fluorescently and coupling analytes to beads which can be distinguished by their intrinsic fluorescence. Presently we label scFvs using a coiled coil approach, in which synthetic fluorescent peptide K coils bind to E coils fused to the scFv. The future use of GFP based binders will eliminate the need for this step. By using specific and non-specific targets, as well as anti-tag antibodies which are able to bind to all binding ligands, we have been able to obtain analyze individual clones from real selections. The information obtained includes binding to specific targets, indications of expression levels, and the degree of polyreactivity/non-specific binding. Depending upon the flow cytometer used, the analysis of each individual affinity reagent clone can be carried out in 1-30 seconds. However, one present difficulty is the export of data, which needs to be carried out manually. With improved data export and analysis, this screening method will be able to handle the throughput desired for the affinity reagents portion of the GTL protein production facility.