[8] Robotic Cloning and Protein Production Platform of the Northeast Structural Genomics Consortium

By Thomas B. Acton, Kristin C. Gunsalus, Rong Xiao, Li Chung Ma, James Aramini, Michael C. Baran, Yi-Wen Chiang, Teresa Climent, Bonnie Cooper, Natalia G. Denissova, Shawn M. Douglas, John K. Everett, Chi Kent Ho, Daphne Macapagal, Paranji K. Rajan, Ritu Shastry, Liang-yu Shih, G.V.T. Swapna, Michael Wilson, Margaret Wu, Mark Gerstein, Masayori Inouye, John F. Hunt, and Gaetano T. Montelione

Abstract

In this chapter we describe the core Protein Production Platform of the Northeast Structural Genomics Consortium (NESG) and outline the strategies used for producing high-quality protein samples using *Escherichia coli* host vectors. The platform is centered on 6X-His affinity-tagged protein constructs, allowing for a similar purification procedure for most targets, and the implementation of high-throughput parallel methods. In most cases, these affinity-purified proteins are sufficiently homogeneous that a single subsequent gel filtration chromatography step is adequate to produce protein preparations that are greater than 98% pure. Using this platform, over 1000 different proteins have been cloned, expressed, and purified in tens of milligram quantities over the last 36-month period (see Summary Statistics for All Targets, http://www.nmr.cabm.rutgers.edu/bioinformatics/ZebaView/). Our experience using a hierarchical multiplex expression and purification strategy, also described in this chapter, has allowed us to achieve success in producing not only protein samples but also many three-dimensional structures. As of December 2004, the NESG Consortium has deposited over 145 new protein structures to the Protein Data Bank (PDB); about two-thirds of these protein samples were produced by the NESG Protein Production Facility described here. The methods described here have proven effective in producing quality samples of both eukaryotic and prokaryotic proteins. These improved robotic and/or parallel cloning, expression, protein production, and biophysical screening technologies will be of broad value to the structural biology, functional proteomics, and structural genomics communities.

0076-6879/05 $35.00

Introduction

The Northeast Structural Genomics Consortium (NESG) is a pilot project designed to evaluate the feasibility and value of structural genomics. Its primary goals are to develop and refine new technologies for high-throughput protein production and structure determination by both NMR and X-ray crystallography and to apply these technologies in determining representative structures of the domain sequence families that constitute eukaryotic proteomes. The project (http://www.nesg.org), one of 11 pilot projects supported by the United States National Institutes of Health Protein Structure Initiative (http://www.nigms.nih.gov/psi/), is developing technology aimed at optimizing each stage of the structure determination pipeline.

One of the most important challenges to the emerging field of structural genomics is the preparation of protein samples suitable for the determination of three-dimensional structures. This sample preparation challenge is different from those encountered in most previous genome-wide initiatives, such as the Human Genome Sequencing Project or microarray gene expression studies, which focus on preparing nucleic acid samples (Lander, 1999; Lander *et al*., 2001; Winzeler *et al*., 1999), or involve production of only small quantities of proteins for functional studies (Ito *et al*., 2001; Uetz *et al*., 2000). Nucleic acids all have generally similar biophysical properties, allowing similar and well-defined purification and preparation techniques to be employed in high-throughput processes. Other genome-wide studies that focus on proteins such as yeast two-hybrid screens (Giot *et al*., 2003; Ito *et al*., 2001; Li *et al*., 2004; Rain *et al*., 2001; Uetz *et al*., 2000) require relatively small amounts of proteins, often expressed in a eukaryotic organism (yeast), and do not usually require protein purification to derive experimental information. Structural genomics projects require the production of tens of milligram quantities of soluble, highly purified protein samples. These proteins often have diverse biophysical properties, making the preparation of suitable samples more difficult, especially when considering high-throughput methods. The target proteins of the NESG (http://www-nmr.cabm.rutgers.edu/bioinformatics/ZebaView/index) are composed of protein domain families sharing structure and sequence similarity selected from the proteomes of archaea, eubacteria, and eukaryotic organisms, many of which are difficult to express in *Escherichia coli* expression systems. In addition, the NESG Consortium utilizes both nuclear magnetic resonance (NMR) and X-ray crystallographic methods of protein structure determination (Montelione and Anderson, 1999). Protein samples suitable for rapid three-dimensional (3D) structure determination by NMR and X-ray crystallography generally require ^{13}C,^{15}N isotope

enrichment or selenomethionine labeling. This necessitates that our protein production platform not only has high throughput but also is flexible enough to handle preparation of both protein sample types. Considering these challenges, one of the major contributions structural genomics will have on science is the development of new technologies that enhance our capabilities in the areas of protein expression and purification, and improve our abilities to deliver protein samples suitable for NMR, X-ray crystallography, and diverse biological studies.

In this chapter, we describe the high-throughput cloning and protein production platform we have developed at Rutgers University for the preparation and screening of protein samples amenable to structural determination by X-ray crystallography and/or NMR spectroscopy. The laboratory of Cheryl Arrowsmith at the Ontario Cancer Institute and the University of Toronto also produces proteins for structure studies by the NESG. This related, though distinct, platform has been described elsewhere (Yee *et al*., 2002, 2003). The process (and most statistics) described in this chapter are specifically for the Rutgers component of the NESG protein production effort.

Although the Rutgers protein production effort for the NESG is currently limited to protein production in *E. coli*, this platform is quite flexible, providing for cloning and expression of a wide range of proteins from archaea, eubacteria, and eukaryotic organisms. The robotic platform is highly efficient, currently providing the capacity to clone and evaluate expression and solubility of 100 proteins per week and to produce tens of milligram quantities of 15–20 purified proteins per week for both NMR and crystallization screening. In addition to its central role in driving our structural genomics effort, the platform is a prototype of protein production technologies that will soon become commonplace in traditional structural biology, biochemistry, and proteomics projects.

Protein Production Platform

Targets and Bioinformatics Infrastructure

Most of the current NESG target proteins are full-length polypeptide chains shorter than 340 amino acids, selected from domain sequence clusters (Liu and Rost, 2004; Liu *et al*., 2004), which are organized in the PEP/CLUP (http://cubic.bioc.columbia.edu/pep/) domain cluster database (Carter *et al*., 2003). Each of these protein sequence clusters consists of three or more proteins (or protein fragments) corresponding to putative structural domains whose 3D structure is not known experimentally and cannot be accurately modeled through homology. The NESG focuses on

TABLE I
EUKARYOTIC TARGET GENOMES OF THE NESG CONSORTIUM

Organism	Number of targets
Arabidopsis thaliana	2242
Caenorhabditis elegans	340
Drosophila melanogaster	263
Homo sapiens	2857
Saccharomyces cerevisiae	584

domain families that include at least one representative from a set of five eukaryotic target organisms (Table I). These correspond to domain families constituting the eukaryotic proteome.

Zeba View (http://www-nmr.cabm.rutgers.edu/bioinformatics/ZebaView/), a web-based interactive summary of key NESG target information, functions as the ''Official Target List'' of the NESG project (Wunderlich *et al*., 2004), and SPINE (Structural Proteomics in the Northeast; http://SPINE.nesg.org/), a web-based project database, organizes and coordinates detailed information about the protein production and structure analysis processes carried out in the multiple sites of the NESG Consortium (Bertone *et al*., 2001; Goh *et al*., 2003). SPINE is a laboratory information management system (LIMS) for most of the steps of the protein production process, as well as a data warehouse of information collected from other laboratory information management systems used by the NESG Consortium through an XML-based data exchange language (Wunderlich *et al*., 2004). Each NESG protein target is assigned a NESG id code, the first letter(s) of which indicate the organism from which the target is cloned, the last letter the institute at which the protein is produced, followed by a serial number (e.g., HR32, human, Rutgers, target number 32).

Multiplex Expression Vector System

Highly homogeneous protein samples with minimal numbers of disordered nonnative residues are generally required for successful protein crystallization and for structure determination by X-ray crystallography or NMR. Protein samples for crystallization should ideally exhibit >98% homogeneity on sodium dodecyl sulfate (SDS) polyacrylamide gels. Moreover, whereas affinity tags are generally required for high-throughput purification protocols (Crowe *et al*., 1994; Sheibani, 1999), large disordered tags can frustrate crystallization efforts and often exhibit strong sharp

peaks and associated artifacts of Fourier transform processing in NMR spectra. Protein samples must be produced in soluble form and at high yield, as tens of milligram quantities are needed for crystallization and NMR experiments. For NMR studies, the high cost of uniform enrichment with ^{13}C isotopes generally demands high efficiency isotope incorporation. For example, in *E. coli* expression systems, we typically aim for production yields of 10–50 mg of purified protein per liter of fermentation using defined minimal media (MJ9) optimized for producing isotopically enriched proteins (Acton *et al.*, 2005; Jansson *et al.*, 1996). These constraints define the primary design features of vectors expressing protein open reading frames (ORFs) for use in structural genomics projects.

The advent of genomic studies has led to the introduction of several new cloning technologies, many of which are optimized for high throughput, including various systems of ligase-independent cloning. These cloning strategies, such as the Gateway (Invitrogen), TOPO (Invitrogen), and Creator (Clontech) systems, exploit various forms of recombinational cloning (Abremski and Hoess, 1984; Hartley *et al.*, 2000; Sauer, 1994; Shuman, 1994). These systems generally exhibit high cloning efficiency and significantly fewer cloning steps, both of which are advantageous for high-throughput procedures. However, as a consequence of the mechanisms of recombinatorial cloning, these strategies generally result in protein products with a significant number of nonnative amino-acid residues attached to one or both ends of the protein molecule. For example, an N-terminal His-tag fusion in the Gateway system results in the addition of 22 extraneous amino-acid residues. These nonnative residues can interfere with crystallization and other structural studies. It is possible to introduce a protease cleavage site downstream of N-terminal tags and thereby cleave off the extraneous residues from the recombination site (Yee *et al.*, 2002, 2003). However, there are often problems with such systems, including protease specificity and/or contaminating proteases that lead to unwanted cleavage(s), incomplete cleavage leading to nonhomogeneous samples or low yields, and the overall cost and complexity that this step adds to the high-throughput pipeline. In addition, many of the most useful sources of cDNA libraries for eukaryotic organisms in Gateway libraries do not include a stop codon allowing both N- and C-terminal fusions to be produced from a single entry clone. This unfortunately adds residues from both recombination sites, producing a protein with a very significant number of nonnative residues. These large numbers of extraneous residues may contribute to the limited success of structural genomics projects using such systems. Moreover, the cleavage pattern of commonly used site-specific proteases, such as the TEV protease (Kapust *et al.*, 2002), leave four to six

residues on the N-terminal side of the recognition site, limiting their usefulness with C-terminal fusion tags.

With these issues in mind, we chose to create an expression vector set that would utilize a classic restriction endonuclease-ligase-dependent mechanism of cloning that could allow the generation of constructs with a minimum number of extraneous residues, while avoiding the requirement for protease cleavage and subsequent purification. Additionally, although a number of different systems for recombinant protein production in bacteria or eukaryotic cells are currently available (Geisse *et al*., 1996; Makrides, 1996), we opted to base our effort on isopropylthiogalactoside (IPTG)-inducible systems already in use for high-level expression in bacteria (Bujard *et al*., 1987; Studier *et al*., 1990), which readily allow isotope and L-selenomethionine (SeMet) enrichment. These pET vector systems also allow use of autoinduction media (Studier, 2004). Our focus was to create a flexible system that could efficiently generate an array of combinations for rapid screening of optimal expression conditions. Because every protein has different properties, which currently cannot be predicted in advance, we wanted the ability to produce the same protein as an N- or C-terminal hexa-histidine (6X-His) fusion (for rapid affinity purification and expression/solubility/NMR screening) as well as a nontagged version (for use in structure determination, if preferable). We also wanted to produce each of these protein variants under a number of different expression conditions (by varying promoters, bacterial strains, etc.), as optimal expression conditions generally vary from one protein to another.

To meet these criteria, we created a ''Multiplex Vector Kit'' consisting of a set of nine compatible expression vectors. The essential features and the minimal polylinker sequences of this vector set are shown in Fig. 1. As a starting point, we used commercially available *E. coli* expression vectors differing in the choice of promoter (T7, T7 lac, or T5 lac lac) and placement of a 6X-His tag at the N- or C-terminus, which we modified to suit our needs. As some of these commercial vectors have very limited polylinkers, we have engineered into these vectors an expanded ''minimal common polylinker'' (MCP) containing a set of restriction endonuclease (RE) sites shared by vectors with more extensive polylinkers. We placed the MCP in all three reading frames (1, 2, and 3) with respect to the 6X-His tag, allowing us to minimize the nonnative residues added to an ORF and to control the identity of amino acid residues that are added between the native sequence and the 6X-His tag. This generated a set of three vectors from each starting vector and also created two new sets of vector cognates that allow the choice of *Nco*I in place of *Nde*I as an option for in-frame ATG cloning (Fig. 1).

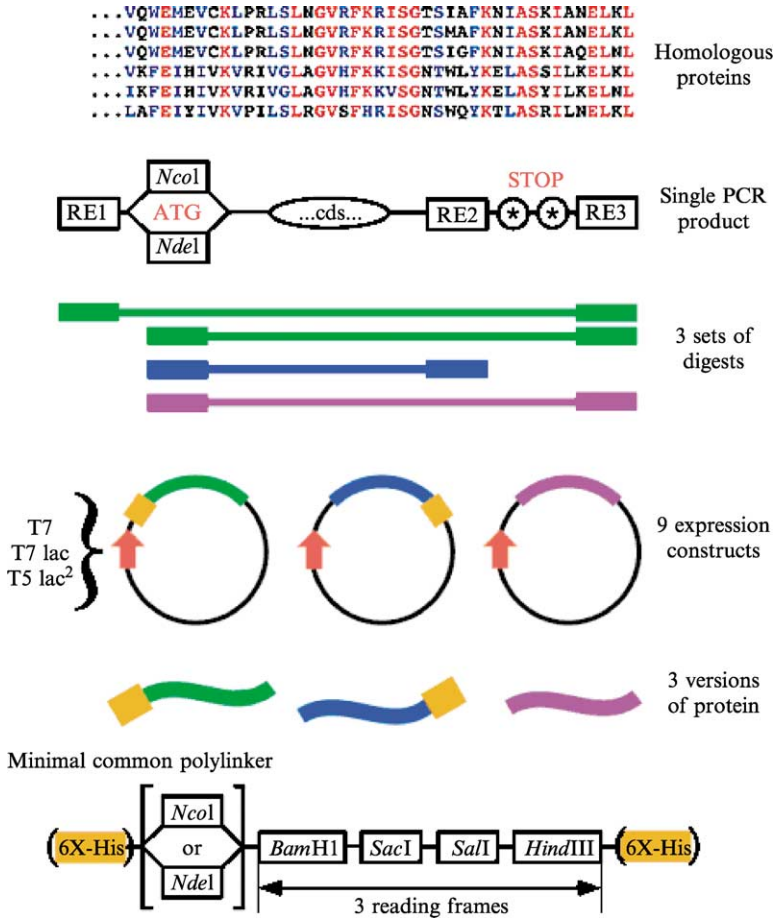


FIG. 1. Strategy for multiplexed protein expression. One or more representatives from a family of homologous proteins or protein domains are chosen. For each domain to be expressed, a PCR product is designed to contain either an *Nco*I or an *Nde*I site at the initiator ATG for the coding sequence along with three additional restriction sites (RE1, RE2, and RE3) from the MCP for cloning different versions of the protein into the various expression vectors. RE1 and either *Nco*I or *Nde*I are included in the 5′ PCR primer, while the 3′ primer includes RE2 followed by one or two stop codons and RE3. Using three different combinations of digests, nine different expression variants can be generated from the same PCR primers, differing in the promoter driving expression, the placement of an affinity tag (if any), and the identity of any nonnative amino-acid residues that result from the cloning strategy. The minimum common polylinker found in each of the nine custom expression vectors is shown at the bottom.

A key goal of our design was to develop a vector set with the flexibility to create the maximum number of different expression constructs using a minimum number of different polymerase chain reaction (PCR) primers, considering that one of the greatest expenses in high-throughput cloning is the cost of oligonucleotide primers and high-fidelity Taq polymerases. With the vector modifications we introduced, a single PCR product can be designed so that only three different double restriction endonuclease digestions are sufficient to generate up to nine different expression constructs in parallel: three protein variants (N- or C-tagged 6X-His fusions, plus the nonfusion) each driven by any of the three promoters (T7, T7 lac, or T5 lac lac). A schematic of this strategy is shown in Fig. 1 (see legend for details). To accommodate all cloning options, PCR primers must be designed to introduce a specific arrangement of restriction sites into the resulting PCR product, which contains the protein coding sequence of interest flanked on either end by restriction sites compatible with the vector polylinkers (RE1, *Nde*I/*Nco*I, RE2, and RE3). A detailed discussion of the essential issues that must be considered in designing the PCR product and cloning with this expression kit is presented elsewhere (Acton *et al.*, 2005; Everett *et al.*, 2004).

We designed the ''Multiplex Vector Kit'' to allow cloning into several (at least nine) different vectors from a single PCR product. However, in the course of our work we have identified a second strategy for which the resulting common polylinker can be used to implement a significant cost-saving advantage. In the first stage of this procedure, we design primer pairs for cloning each of a large set of ORFs into one of the C-terminal 6X-His vectors, which allows a decreased number of nucleotides per primer as only one six-base restriction site is added. These PCR products are then cloned into one of the C-terminal fusion vectors from the Multiplex Vector Kit. In the second stage, each of the resulting C-terminal fusion constructs can then be used as a PCR template for amplification using a set of primers that anneals to the vector sequence flanking the ORF while introducing a stop codon and a restriction site at the 3′ end of the gene (Fig. 2). This is accomplished using a 3′ primer that anneals to the 6X-His coding sequence, bubbling off a restriction site and a stop codon, and then annealing back to the 3′ restriction site into which the gene was originally cloned. The resulting PCR product is then cloned into an N-terminal 6X-His or nontagged vector of the Multiplex Vector Kit using the original 5′ RE site together with the newly added 3′ RE. This adds only two nonnative residues derived from the original 3′ RE site that is now directly followed by the introduced stop codon. In this manner, all nine different expression constructs can be derived with a minimal number of initial primers and the cost-effective common primers.

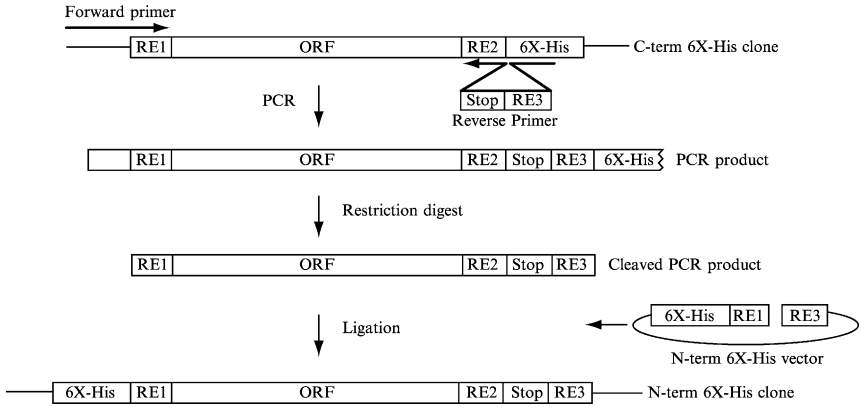


FIG. 2. Using common primers for shuffling targets between multiplex vectors. This schematic illustrates a strategy for producing all other construct variants in the "Multiplex Vector Kit" from a single C-terminal 6X-His construct using primers that are not gene specific. The forward primer anneals to the common vector sequence upstream of the coding sequence and will incorporate the initial RE1 site. The reverse primer anneals to the 6X-His coding sequence and includes a new restriction site (RE3), present in the other vectors, as well as a TAG stop codon. The primer then anneals to the original RE2 site, directing elongation from this point. The resulting PCR product is cleaved with RE1 and RE3, removing the common vector sequence and the original 6X-His coding sequence, and then ligated into a similarly cleaved vector. In this illustration, the ORF is then cloned into an N-terminal 6X-His fusion vector; this strategy can also be used to produce nontagged variants.

RT-PCR for High-Throughput Cloning of Eukaryotic ORFs

Using the strategy described above, high-throughput cloning of structural genomics targets from prokaryotic genomes is relatively straightforward. In particular, the absence of introns allows for the direct use of genomic DNA as a PCR template. As a consequence, although the primer sets for each ORF differ, the DNA template is common to all of the PCR reactions. This allows for simple manual or robotic manipulation using a common PCR cocktail containing all of the required buffers, enzymes, and the genomic template. However, a bottleneck emerges when cloning from most eukaryotic target genomes, since many of these genes contain introns. This necessitates using cDNA as a PCR template, resulting in several complications. Most importantly, adequate cDNA libraries must be obtained. The highest quality full-length libraries will be those arising from large-scale projects such as the Drosophila Gene Collection, I.M.A.G.E., or the *Caenorhabditis elegans* ORFeome project (Reboul *et al.*, 2003; Rubin *et al.*, 2000; Stapleton *et al.*, 2002; Strausberg *et al.*, 1999, 2002). A structural genomics project focusing on thousands of eukaryotic targets

requires thousands of cDNA clones to serve as PCR templates; handling this set of individual cDNA clones incurs significant logistical complications and additional costs. For example, in genomic-scale cloning it is most practical to acquire the entire gene set, which is not only costly but also requires sufficient resources for archiving and retrieving the reagents. In addition, the use of target-specific cDNA templates complicates robotic automation, since each individual template must be transferred to an appropriate well in a PCR plate, presenting additional bioinformatics, robotic programming, and material costs. The resulting increased complexity also lengthens the time required for setting up the PCR reactions and generally has a negative effect on the outcome of the amplification.

To circumvent the problems associated with using target-specific cDNA templates and to increase throughput, we instituted a reverse transcriptase strategy to produce a common cDNA pool for use as a PCR template. In this strategy, we use polyadenylated mRNA or total RNA from various tissues, cell types, and developmental stages together with oligo(dT) primers to carry out reverse transcriptase reactions. Briefly, oligo($dT_{12–18}$) (Invitrogen) is annealed to 5 μg of RNA in a volume of 275 μl by heating to 70° for 10 min followed by incubation on ice for 15 min. The volume is raised to 500 μl with the addition of Powerscript Reverse Transcriptase (Clontech), the corresponding first strand synthesis buffer, free dNTPs, and RNase-free water. The reaction is incubated for 60 min at 42° allowing first strand cDNA synthesis to occur, followed by digestion with RNase H (New England Biolabs) ensuring the removal of RNA that might interfere with PCR amplification of our target sequences (for greater detail see Acton *et al.*, 2005). For each organism, cDNAs from several tissues, cell types, and/or developmental stages are then mixed to form a common cDNA pool. This cDNA pool is then added to the PCR cocktail mix, much like adding bacterial genomic DNA, and used as a common cDNA template in PCR reactions.

Robotic Vector Construction with the Biorobot 8000

To clone in a high-throughput manner we have automated each step of our restriction endonuclease-ligase-dependent cloning strategy using a Biorobot 8000 (Qiagen). Figure 3 outlines each of these steps of vector construction; steps shown in blue typeface are completely automated by the robot, while those in red are semiautomated, requiring some manual manipulations. A detailed description of the entire process is provided in Acton *et al.* (2005). Briefly, 96 protein targets are chosen for cloning and the primer pair for each ORF is determined using the Primer Prim'er oligonucleotide design program (Everett *et al.*, 2004). The

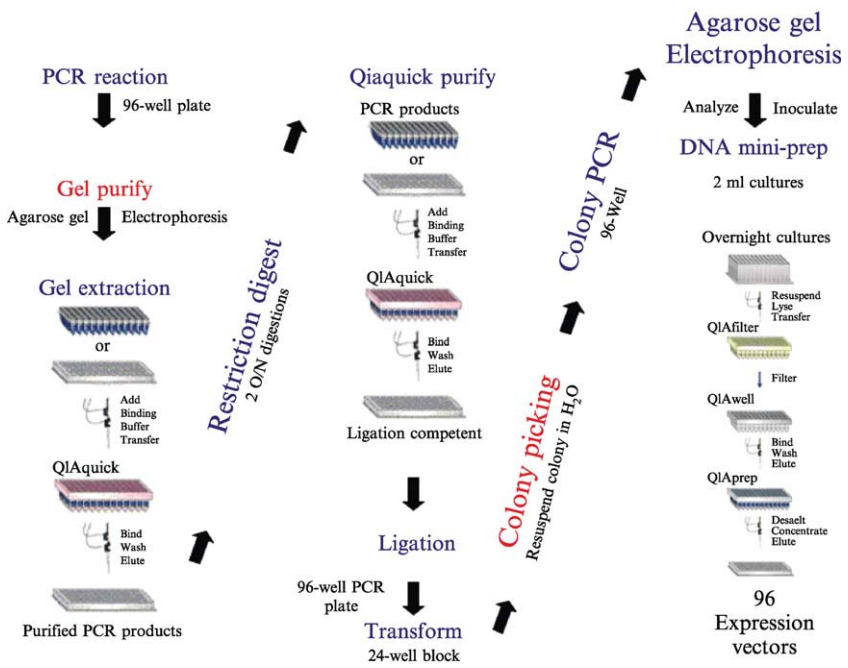


FIG. 3. Biorobot 8000 cloning schematic. Each key step in the cloning strategy is indicated; blue type denotes those steps that are completely automated, and red type indicates those steps that require some manual input. Roughly 1 week of one full-time equivalent is needed to complete all of the cloning steps for 96 target proteins. Several of the procedures are modifications of Qiagen-based protocols, such as the Qiaquick Purification and the DNA Mini-Prep protocols. However, most have been completely created in the Rutgers NESG Protein Production laboratory. A more detailed description of the cloning procedure, as well as the automated protocols, are provided elsewhere (Acton *et al.*, 2005).

Primer Prim'er program (available on-line at http://www-nmr.cabm.rutgers.edu/bioinformatics/) generates order forms for the primer sets, which are transmitted directly to the primer vendor, typically Qiagen. Forward and reverse primers are grouped together, with the forward and reverse primers for each specific ORF in identical well positions on two separate order forms, synthesized by Qiagen, and provided in 96-well format with the concentration of each primer normalized to 50 μM in deionized water. The two 96-well primer blocks containing the reverse or forward primers are placed on the robot deck, and a Qiasoft 4.0 program written to automate the PCR setup is run. In this program, a PCR reaction mix, containing all necessary components for PCR amplification, is added to each well in a 96-well PCR plate. This includes the dNTPs, Advantage HF2 high-fidelity polymerase (Clontech), and its corresponding buffer and TaqStart

antibody (Clontech). The latter sequesters the polymerase prior to thermocycling, which we have found to greatly decrease background amplification products caused from mispriming during the low temperature PCR setup, while also increasing the yield of correctly amplified products. The program then commands the pipette head to transfer 100 pmol of the appropriate forward and reverse primers from the primer blocks into the corresponding well for each target in the PCR plate. An Applied Biosystems 9700 thermocycler is used for the amplification with 35 total cycles. Each cycle contains a 10-s 90° melting step, a 30-s annealing step (50–55°), and a 1-min 68° elongation step. An annealing temperature step increase after 10 rounds of amplification is included to take into account the contribution of the extra bases added for the restriction sites (for greater detail, see Acton *et al*., 2005).

Following PCR amplification, the products are separated on a 2% agarose gel, and the DNA bands are visualized using a low-energy ultraviolet (UV) lightwand. The correct-size fragments are easily identified, since the primer design program organizes the ORFs in the plate by increasing size (Everett *et al*., 2004). The proper DNA fragments are then manually excised from the gel using a scalpel and relocated into the appropriate well of a 96-well S-Block (Qiagen). A completely automated 96-well gel extraction is carried out using reagents from the Qiagen Gel Extraction Kit and a QIAquick 96-well column PCR Cleanup plate. The resulting purified PCR products are then subjected to two restriction endonuclease digestions to allow for directional cloning, generally using *Nde*I and *Xho*I at the 5′ and 3′ ends, respectively. Following the second restriction digestion, an automated 96-well Mini-Elute DNA purification and elution into water is performed. Ligation into an appropriate precut expression vector is then carried out. Briefly, a 96-well PCR plate is chilled on the robot deck and a reaction mixture containing 100 ng of a similarly digested vector, ligase buffer, ligase (100 U, New England Biolabs), and water is trans erred to each well. Three- to 6- fold-molar excess (generally 1 or 2 μl) of the highly purified and cleaved DNA PCR product is added to the appropriate well for a 20-μl final volume, mixed, and incubated overnight at 16°.

Having completed vector construction, the next step of the process involves robotic transformation into *E. coli* cells in 24-well format. A 1-μl aliquot of each overnight ligation well is pipetted by the Biorobot 8000 into a corresponding well in another 96-well PCR plate prechilled at 0° on the robot deck. Each well of this plate contains 15 μl of XL-1 ultracompetent cells (Stratagene). A transformation procedure is then carried out on the robot deck keeping the PCR plate at 0° until a manual heat shock. SOC (100 μl) is added to each well, and the plate is incubated at 37° for 1 h.

The transformation is completed by pipetting the entire contents of each well into the corresponding wells of four 24-well blocks. Each block well contains 2 ml of Luria broth (LB) medium/Agar with ampicillin and 5–10 (3-mm-diameter) glass beads. The contents are dispersed using the robot's platform shaker and the glass beads, the latter of which are then poured off the plate. Following overnight incubation at 37°, two colonies per ORF are harvested and resuspended in 50 μl of sterile water. Colony-picking is the most labor-intensive step in the process outlined in Fig. 3. Colony PCR, using primers flanking the MCS, is set up robotically in 96-well format, and the results are visualized by agarose gel electrophoresis, identifying clones with correct-size inserts. These clones are then subcultured overnight, and plasmid DNA is isolated using a completely automated Qiagen 96-well DNA mini-prep procedure.

Archiving Expression Vectors

Considerable time, effort, and funds have gone into making each expression constructs, and during the structural determination process it is often necessary to produce multiple large-scale protein preparations. It is of the utmost importance that each expression construct is archived in a manner sufficient to allow easy retrieval and secure storage. Before each construct is miniprepped, two glycerol stocks in 96-well format are produced by aliquoting from the overnight culture and adding glycerol to a final concentration of 20% followed by flash freezing in dry ice. In addition, after the DNA miniprep is completed, an aliquot of each new construct is added to the appropriate well of two new 96-well plates that are then lyophilized to dryness, while the original is kept in liquid form as a working stock. Both the glycerol stocks and the DNA plates are then stored at −80°, with duplicates residing in separate freezers. The position of each plate and the contents of each well are then uploaded into the SPINE database such that each construct record in SPINE has associated DNA and glycerol stock locations. In this manner, the location of each clone in either form can be quickly located, using the Web-based SPINE LIMS, and subsequently retrieved.

Robotic Protein Expression Screening with the Biorobot 8000

The large number of expression constructs created by the automation of cloning also necessitates a large capacity screening process to evaluate the efficiency of protein expression and protein solubility in a high-throughput manner. Although all of the expression constructs could potentially be screened on a preparative scale, this would be costly and inefficient, since a large fraction of targeted proteins is observed to be

either insoluble or not expressed in *E. coli*. The goal of small-scale expression is to predict the expression and solubility of each construct on the preparative scale. It should therefore be as representative of the large-scale conditions as possible. Moreover, to the degree that the analytical scale screening results correlate with large scale expression results, the smaller scale experiments can be used to explore different expression conditions, such as alternate bacterial strains, since different conditions sometimes produce significantly different expression or solubility results.

The scheme in Fig. 4 outlines our robotic 96-well expression and solubility screening process. Similar to the cloning schematic, completely automated steps are shown in blue and the partially automated steps are in red. Briefly, the starting material for the expression screening is the miniprep DNA derived from the cloning steps. Although it is possible to clone directly into an expression strain, we prefer to initially transform into a more stable cloning strain for archival purposes and then perform fresh transformations as needed in appropriate expression strains. We generally use the codon-enhanced BL21 (DE3)pMgK strain, containing plasmid-derived genes for arginine and isoleucine tRNA, since the codon usage in bacteria can be quite different than in eukaryotic organisms resulting in poor translation (Chen and Inouye, 1990; Ikemura, 1985; Sorensen *et al.*, 1989). After the completely automated transformation, individual colonies are picked from the four 24-well plates and inoculated into the corresponding well of a 96-well block (2.2 ml) containing 0.5 ml of LB medium per well. This initial culture is grown for 6 h at 37°, and a small aliquot from each well is added by the Biorobot 8000 to the corresponding well of a fresh 96-well block containing 0.5 ml of MJ9 minimal media (Jansson *et al.*, 1996) for overnight growth. Following saturated growth, the robot performs a 1:20 dilution into the corresponding well of one of four 24-square-well blocks (10 ml maximum volume/well) containing 2 ml of MJ9 media (Jansson *et al.*, 1996), covered with Airpore tape (Qiagen), and grown to mid-log phase (2–3 h growth) with vigorous shaking at 37°. The small volume of media in conjunction with the gas-permeable tape allows for excellent aeration, similar to the baffled Furnbach flasks used for large-scale protein synthesis, allowing the results of our analytical expression testing to more accurately mirror the results of subsequent preparative-scale fermentations. Once mid-log phase (0.5–1.0 OD_{600} units) has been reached, determined by sampling several wells in each plate, expression is induced with IPTG, the temperature is shifted to 17°, and the cultures are grown overnight with vigorous shaking. It has been previously reported, and we have also observed, that low temperature induction is often helpful in aiding solubility (Shirano and Shibata, 1990). Cells are harvested by centrifugation, the pellets are resuspended in lysis buffer

Transformation
96-well plate
BL21
24-well block
96-well culture
2.2 ml S-Block
37°
LB
Overnight culture
96-well transfer (MJ9)
37°
2.2 ml S-Block
24-well blocks
96 to 24-well transfer
37°
Induction
OD600 ~0.4 – 0.8
IPTG
O/N @17°
Harvest
Plate centrifuge
Plate sonicate
96-well plate
96-well Ni-NTA
Turbofilter
Clear lysates
Vacuum
QIAquick-Ni NTA superflow
Bind
Channeling block
Vacuum
Wash
Channeling block
Vacuum
Elute
Vacuum
Purified protein
96-well plate
Transfer protein
Add reagent
Bradford
Plate reader
96-well plate
Concentration
Well Abs ug/ml
A6 0.477 137
B8 0.516 152
C7 0.279 60
E6 0.771 251
E9 0.37 95
F4 0.359 91
G4 0.66 208
D10 0.092 <1
B4 0.201 30
A6 B8 C7 E6 E9 F4 G4 D10 B4
SDS-PAGE

(50 m*M* NaH_2PO_4, 300 m*M* NaCl, 10 m*M* 2-mercaptoethanol) and robotically transferred to a 96-well PCR plate. A 96-well sonicator is used to break open the cells, the lysates are added directly to the robot platform, and the His-tagged recombinant proteins are purified using a modified 96-well Ni-NTA purification protocol (Qiagen).

An aliquot of each well is next transferred to a fresh 96-well plate containing Bradford reagent (Bradford, 1976) and the absorbance measured using a plate reader (see Fig. 4). The concentration of soluble, expressed protein competent for Ni binding is automatically calculated from the absorbance, and constructs returning greater than a calculated 5 mg of protein per liter of culture are marked for large-scale expression and purification. The SDS–polyacrylamide gel electrophoresis (PAGE) gel in Fig. 4 shows a representative sample of proteins purified in this manner, together with data demonstrating the good correlation between protein concentration estimates by automated Bradford and Comassie Blue band stain intensity. These data are archived in the SPINE database and used to identify constructs providing good protein expression and solubility for scale up and biophysical analysis.

Fermentation and Preparative-Scale Protein Expression

Although the analytical expression analysis is invaluable in ascertaining the behavior of each target when expressed in bacteria, the amount of protein provided (10–500 μg) is not sufficient for crystallization experiments or structure determination. Therefore, the expression process needs to be scaled up such that 10–100 mg of purified protein can be produced, necessitating larger culture volumes. Our process for preparative-scale protein expression, shown in Fig. 5, has been designed to optimize conditions with respect to yield, cost, throughput, and the different structural determination approaches. We opted against using 1 liter or small fed-batch fermenters mainly for reasons of cost, both of equipment and

FIG. 4. High-throughput analytical scale protein expression screening using robotic methods. This schematic shows the step-by-step procedure used for small-scale expression screening. Completely automated steps are shown in blue, and red denotes steps that are partially automated. The entire process is conducted in 96-well plates or a corresponding number of 24-well blocks. The right top shows a modified 96-well Bradford assay (Bradford, 1976) with aliquots from the 96-well Ni-NTA purification. The plate configuration is the normal 8 rows by 12 columns. More intense blue wells denote a higher concentration of purified protein and hence constructs that express high levels of soluble proteins. These targets are slated for large-scale production. The relative concentration is calculated by the 96-well plate reader and is reported in spreadsheet format (see the blue box). An SDS–PAGE gel shows the results of the purification and relative agreement with the calculated values.

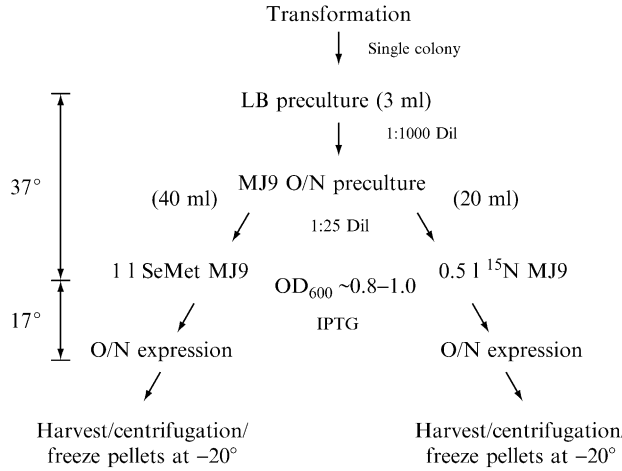


FIG. 5. Preparative-scale protein expression. Schematic of protein expression for NMR and X-ray crystallography samples. Each target is transformed into an appropriate BL21 (DE3) strain and subcultured into minimal media. Each target diverges into two pathways, for isotope enrichment and selenomethionine labeling, respectively. Preliminary growth occurs at 37° and the temperature is shifted to 17° upon IPTG induction. O/N, overnight.

reagents, as well as the fact that the prohibitive cost limits parallelization. Therefore, a strategy based on growth in 2-liter baffled Furnbach flasks was chosen, based on the simplicity of the technique, the low cost of the required equipment, and the ease of parallelization. Though not used in our platform, it is also possible to utilize disposable 2-liter plastic bottles in place of these Furnbach flasks (Millard *et al*., 2003). Having decided on this method, the growth conditions were optimized to provide high yields while maintaining ease and throughput.

The growth medium for protein production is MJ9, a modified minimal medium containing a stronger buffering system and supplemental vitamins and trace elements (Jansson *et al*., 1996), which has been optimized for efficient isotopic enrichment of proteins. We have found that MJ9 medium can support the same cell density and protein expression levels as rich media such as LB (data not shown), although not as high as superrich media such as Terrific Broth (Tartof and Hobbs, 1987). NMR studies of proteins generally require enrichment with ^{15}N, ^{13}C, and/or ^{2}H isotopes, using minimal media in which the sole sources of carbon (glucose) and nitrogen (ammonium ion) are uniformly enriched with ^{13}C and ^{15}N, respectively. In the absence of a structural model suitable for applying molecular replacement methods, high-throughput X-ray crystallography of protein structures is most efficient using single (SAD) and multiple anomalous diffraction (MAD) methods

(Dodson, 2003; Hendrickson and Ogata, 1997), which are generally readily carried out with SeMet substituted protein samples (Doublie *et al*., 1996; Hendrickson and Ogata, 1997; Hendrickson *et al*., 1990). Both isotopic ^{15}N, ^{13}C, and/or ^{2}H enrichment and SeMet labeling are carried out in our platform using MJ9 minimal media (Jansson *et al*., 1996).

As shown in Fig. 5, fermentation for protein sample production is split into two branches, based on our need to produce proteins for NMR and X-ray analysis with their isotope or amino acid derivatives. The process begins with transformation of the target expression vector into the appropriate BL21(DE3) strain of *E. coli*, followed by an LB preculture. This preculture is then used to inoculate two overnight cultures (20 and 40 ml for ^{15}N and SeMet incorporation, respectively), which are grown to saturation. The entire volumes of each overnight culture are then used to inoculate each of two 2-liter baffled flasks per target, one containing 0.5 liter of MJ9 supplemented with uniformly (*U*)-$^{15}NH_4$ salts (1–2 g/liter) as the sole source of nitrogen and the other with 1 liter of MJ9 containing SeMet (L-selenomethionine at 60 mg/liter). When SeMet is included in the media, cells down-regulate the synthesis of methionine and incorporate the SeMet into nascent proteins (Doublie *et al*., 1996). The cultures are incubated at 37° until OD_{600} ~0.8–1.0 units, equilibrated to 17°, and induced with IPTG (1 m*M* final concentration). Incubation with vigorous shaking in a 17° room continues overnight followed by harvesting through centrifugation. Aliquots of the induced cells are taken and SDS–PAGE analysis is performed on sonicated samples to assay for expression and solubility. The cell pellets, an isotope-enriched sample, and a SeMet-containing sample are generated for each target in this manner, and then stored at −20° until called for through the SPINE information management system by the protein purification team. To maintain cost-effectiveness (as ^{13}C enrichment is considerably more expensive than ^{15}N enrichment), the initial isotope-enriched sample is produced with ^{15}N enrichment only. If NMR screening results on this sample (described below) indicate that the protein is amenable to structural determination by NMR, additional protein samples are prepared with *U*-^{15}N,^{13}C enrichment (and sometimes also partial or complete ^{2}H enrichment) for 3D structure determinations.

Protein Purification

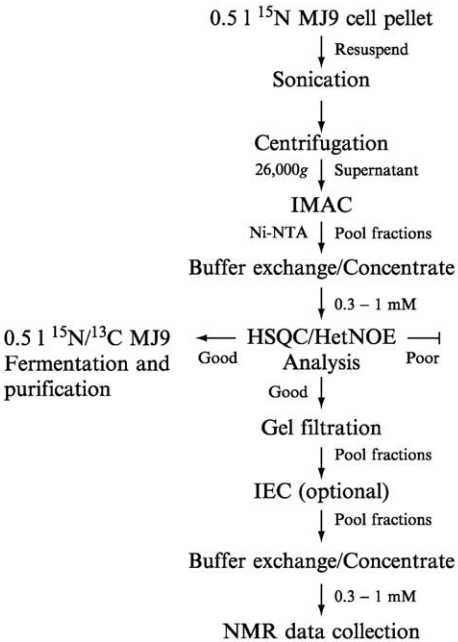
For crystallization and structural studies, it is imperative that the protein samples are highly homogeneous. The need to produce protein samples of sufficient purity while retaining high throughput is the primary challenge of this section of the pipeline. This is especially significant when considering the fact that proteins have such diverse biophysical characteristics.

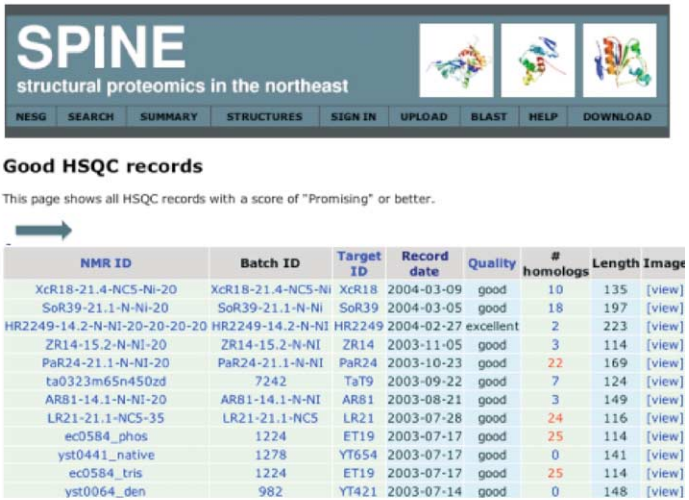
The first step in allowing high-throughput handling is the addition of the 6X-His affinity tag to impart a similar chromatographic characteristic to all of the proteins, thus allowing a common purification technique [immobilized metal affinity chromatography (IMAC)] to be used for all samples, parallelization, and thus high-throughput handling (Crowe *et al.*, 1994; Sheibani, 1999). Like the fermentation pipeline, our protein purification strategy is also divided into two branches, producing samples for NMR (Fig. 6) and X-ray crystallography studies (Fig. 7). In both cases, cell pellets are resuspended in lysis buffer (defined above) with 10 m*M* imidazole (Sigma), lysed by sonication, and centrifuged to pellet the insoluble portion. The resultant supernatant is then applied to nickel-charged HiTrap fast protein liquid chromatography (FPLC) columns (Pharmacia) or nickel-nitrilotriacetic acid (Ni-NTA) agarose (Qiagen) open columns. The loaded columns are then washed with lysis buffer in two steps containing increasing amounts of imidazole, and finally eluted with lysis buffer containing 250 m*M* imidazole. Previously we utilized an AKTAexplorer 3D system running six HisTrap columns sequentially for automated purification of up to six targets. However, this was a timely process that lacked the robustness to handle our high-throughput needs. We have now incorporated a four-module AKTAxpress system and an automated two-step purification using 16 Hi-trap Ni columns and four HiPrep 26/10 desalting columns. This strategy allows for the purification and buffer exchange of 16 target proteins in less than 12 h. This system also has the ability to perform 16 Ni-affinity purifications and gel filtration chromatography steps in an automated fashion and has thus far proven extremely robust.

The path for NMR and X-ray samples diverge at this point in the purification scheme (cf. Figs. 5 and 6). The preparations slated for NMR are sufficiently pure (80–90%) at this step for NMR screening, as described below. IMAC purified SeMet-labeled proteins destined for crystallization screening are next concentrated by ultrafiltation to ~10 mg/ml, exchanged into storage buffer [50 m*M* Na_2HPO_4, 10 m*M* D,L-dithiothreitol (DTT), 300 m*M* NaCl, 100 m*M* arginine, 250 m*M* imidazole, 5 m*M* 2-mercaptoethanol, and 10% glycerol (pH 8.0)], flash-frozen in aliquots, and stored at −80° until prepared for aggregation screening or preparative gel filtration chromatography.

NMR Screening of Ni-NTA-Purified Samples

Although the spectra of many of the targets can be improved with further purification, overall the general amenability to structural determination by NMR can be ascertained with the IMAC purified preparations. Briefly, the protein preparation is divided into three fractions. Each





NMR ID	Batch ID	Target ID	Record date	Quality	# homologs	Length	Image
XcR18-21.4-NC5-Ni-20	XcR18-21.4-NC5-Ni	XcR18	2004-03-09	good	10	135	[view]
SoR39-21.1-N-Ni-20	SoR39-21.1-N-Ni	SoR39	2004-03-05	good	18	197	[view]
HR2249-14.2-N-NI-20-20-20-20	HR2249-14.2-N-NI	HR2249	2004-02-27	excellent	2	223	[view]
ZR14-15.2-N-NI-20	ZR14-15.2-N-NI	ZR14	2003-11-05	good	3	114	[view]
PaR24-21.1-N-NI-20	PaR24-21.1-N-NI	PaR24	2003-10-23	good	22	169	[view]
ta0323m65n450zd	7242	TaT9	2003-09-22	good	7	124	[view]
AR81-14.1-N-NI-20	AR81-14.1-N-NI	AR81	2003-08-21	good	3	149	[view]
LR21-21.1-NC5-35	LR21-21.1-NC5	LR21	2003-07-28	good	24	116	[view]
ec0584_phos	1224	ET19	2003-07-17	good	25	114	[view]
yst0441_native	1278	YT654	2003-07-17	good	0	141	[view]
ec0584_tris	1224	ET19	2003-07-17	good	25	114	[view]
yst0064_den	982	YT421	2003-07-14	good	0	148	[view]

FIG. 6. (*Left*) Protein purification for NMR screening. Isotope-enriched cell pellets are resuspended in lysis buffer, sonicated, and cleared by centrifugation. Following Ni-NTA (Qiagen) IMAC purification, protein-containing fractions are pooled, concentrated, and exchanged into three buffers, which vary pH among other components. HSQC and HetNOE analysis is performed on these samples. If ''good'' spectra are obtained, further purification (gel filtration and optionally ion-exchange chromatography) is performed. (*Right*) View of ''Good HSQC'' Summary Page from the SPINE Database. This page lists those samples that are amenable to structural determination by NMR. Important aspects of the interface include the ability to view an image of the two-dimensional ^{15}N-^{1}H-HSQC for the listed target by selecting ''[view]'' under the image column. In addition, the number found in the ''# homologs'' column indicates how many additional protein targets from this Rost cluster family are in the NESG target list; selecting this link provides a list of these homologs and the progress by the consortium on each member of the family.

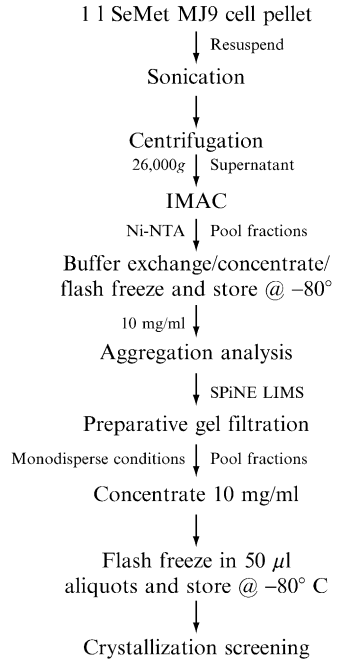


FIG. 7. Protein purification for crystallization screening. Cell pellets from selenomethionine-labeled protein fermentations are resuspended in lysis buffer, sonicated, and cleared by centrifugation. IMAC purification is performed and fractions containing the protein of interest are exchanged into storage buffer. Aliquots are then exchanged into a series of test buffers, and the aggregation state of the protein is assayed by analytical gel filtration and static light scattering (see Fig. 7). Once buffer conditions favoring a single stable species (e.g., monomer, dimer, etc.) are reported, preparative gel filtration in the corresponding buffer is performed. Protein samples are then concentrated to ~10 mg/ml and the preparation is divided into 50-μl aliquots and flash frozen. These samples are then used for high-throughput crystallization screening.

fraction is exchanged into one of three NMR sample buffers (Table II) that differ in pH (4.5, 5.5, and 6.5, including only pH values different from the p*I* of the protein), using dialysis cassettes. This acts as the first step in sample optimization for NMR data collection. The dialyzed samples are concentrated to 0.3–1 m*M* in a final volume of 500 μl, transferred to 5-mm NMR tubes (Wilmad, 535PP), and stored at 4° until NMR data are collected. The sample description, including protein concentration and buffer conditions, is then entered into the SPINS NMR database (Baran *et al.*, 2002), and a subset of this information is transferred automatically

TABLE II
NMR SCREENING BUFFERS

pH	Buffer
6.5 ± 0.1	20 m*M* MES, 100 m*M* NaCl, 5 m*M* $CaCl_2$, 10 m*M* DTT, 0.02% sodium azide, 5% D_2O
5.5 ± 0.1	20 m*M* NaOAc, 100 m*M* NaCl, 5 m*M* $CaCl_2$, 10 m*M* DTT, 0.02% sodium azide, 5% D_2O
4.5 ± 0.1	20 m*M* NaOAc, 100 m*M* NaCl, 5 m*M* $CaCl_2$, 10 m*M* DTT, 0.02% sodium azide, 5% D_2O

into the central SPINE database using an XML exchange language (see Wunderlich *et al.*, 2004, for a description of our basic XML exchange dictionary). SPINS also provides a Web-based list of all targets ready for NMR screening and archives key experimental data and data collection parameters from the NMR instrument.

NMR screening is performed using 500 or 600 MHz NMR spectrometers and is divided into two major components, with each target characterized in several (typically the three pHs described above) buffer conditions. Screening records two-dimensional ^{15}N–^{1}H heteronuclear single-quantum coherence (HSQC) and two-dimensional ^{15}N–^{1}H heteronuclear Overhauser effect (HetNOE) spectra, both usually performed at 20°. The spectral dispersion, together with the relative number of negative-valued peaks in the HetNOE spectrum, quickly indicates if the protein is largely folded; samples exhibiting minimal spectral dispersion and large numbers of negative HetNOE peaks are scored as ''unfolded.'' Samples exhibiting very broad or relatively few peaks, and which are not characterized as ''unfolded'' by HetNOE data, are scored as ''poor.'' A score of ''unfolded'' or ''poor'' indicates that these target samples are not amenable to structure determination by NMR. Samples providing well-resolved and fairly complete HSQC spectra are scored for their amenability to structural determination by NMR, subjectively rated as ''excellent,'' ''good,'' or ''promising'' (Yee *et al.*, 2002, 2003), based on the dispersion of resonances and the percentage of expected peaks (defined by the primary sequence) detected. This information, together with the raw free-induction decay (FID) data and a representative 2D plot of the spectrum, is archived into the Standardized Protein NMR Data Storage and Analysis System (SPINS) database (Baran *et al.*, 2004). These data are then transferred automatically from SPINS to the SPINE database, which is accessible over the internet to the entire NESG Consortium.

Protein samples with an NMR screening score of good or excellent are amenable to structural determination by NMR. NESG researchers are quickly informed of these targets through a ''Good HSQC'' table generated by SPINE (Fig. 6, right). Based on the results in SPINE, researchers then select targets with good or excellent HSQC spectra for pursuit. Following email notification that a target has been selected for NMR structure determination, the original ^{15}N-enriched protein sample is then further purified by gel filtration chromatography and ion-exchange chromatography (the latter only if the gel filtration–purified sample is not sufficiently pure) and concentrated to 0.3–1.0 m*M* in an optimized buffer (as indicated by the ''button test,'' described in the section '' 'Button Tests' to Optimize Condition for NMR Studies'') using ultrafiltration, producing initial samples ready for production data collection. In addition, once selected for structural determination through the ''Good HSQC'' table, the protein target is also scheduled for refermentation in ^{15}N,^{13}C-enriched minimal media and production of a fully double-enriched sample.

Aggregation Screening of Ni-NTA-Purified Protein Samples

It is now well established that proteins that are monodisperse in solution are more likely to produce crystals during screening trials than polydisperse or aggregated samples (Ferre-D'Amare and Burley, 1994, 1997; Manor *et al*., 2005). In an effort to increase the number of samples that produce crystals, the NESG has developed a system that measures the aggregation state of protein samples following gel filtration FPLC, using a combination of static light scattering and refractive index (Manor *et al*., 2005). In this system, analytical gel filtration is carried out using an Agilent 1100 liquid chromatography system with a Shodex Protein KW-802.5 size-exclusion column. The effluent is detected using (1) static light scattering at three angles (45°, 90°, and 135°) measured with a miniDawn (or Dawn) static light-scattering system (Wyatt Technology), (2) absorbance at 280 nm, and (3) refractive index using an Optilab Interferometric Refractometer (Wyatt Technology). Analysis of these data provides estimates of shape-independent weight-average molecular mass (MW_w) and characteristics of the biopolymer mass distributions.

For polydisperse samples, a significant percentage of the mass injected into the system is distributed in multiple elution species, whereas for monodisperse samples the vast majority (>90%) of the mass injected elutes as a single species (e.g., all monomer, all dimer, etc.). Key data from these analyses are archived in the SPINE LIMS. Representative data from an aggregation screen analysis is shown from the corresponding SPINE view in Fig. 8. Using this system, buffer conditions (salt conditions and other

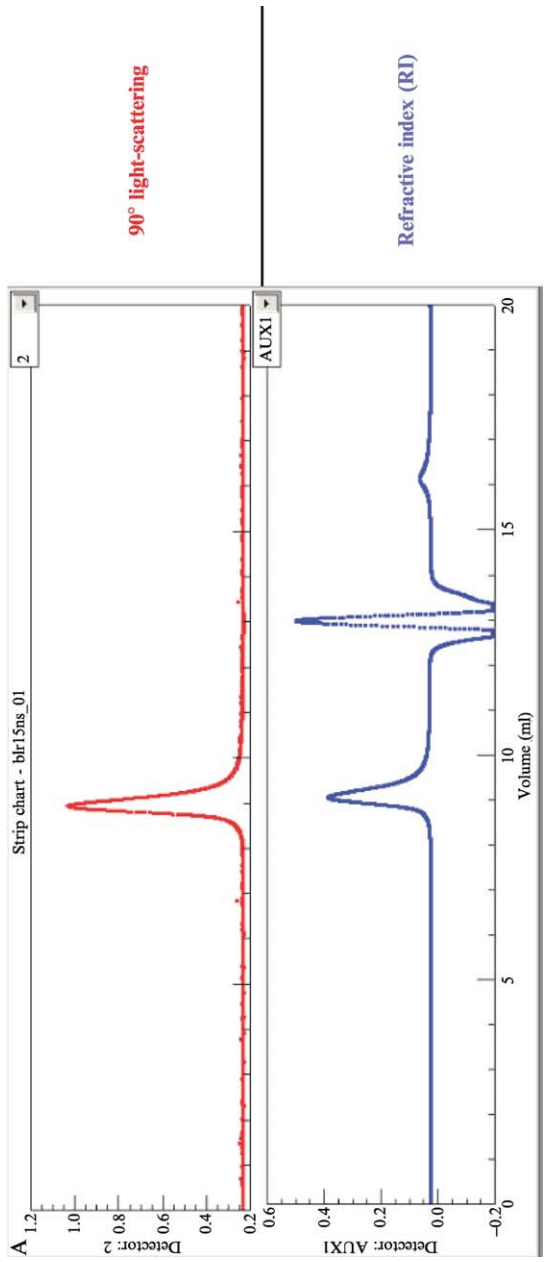


FIG. 8. *(continued)*

B

Aggregation Screening Record for blr15

Recommended Buffers :

- No Salt: 10mM Tris, 5mM DTT

Aggregation Screening 3 EDIT

Prep: BlR15.005NS **Batch:** BlR15-21.1-SeM-Ni	**Researchers:** Bonnie Cooper	
Storage Buffer: 10mM Tris, 5mM DTT		
Storage Buffer Recommended/Requested?: Y		
Exchange Date: 2004-05-26	**Concentration Date:** 2004-05-27	**Analysis Date:** 2004-06-01
Storage Temperature: 4 ° C	**Storage Time:** 5 days	**# F/T Cycles:** 0
Analysis Temperature: 23 ° C	**Analysis Column:** Shodex KW-802.5	**Detector Temperature:** 23 ° C
Analysis Buffer: 250ppm Azide, 100mM NaCl, pH 100		
Protein Stock Volume Consumed: 0.5 mL	**Total Sample Volume :** 1 mL	**Nominal Sample Concentration:** 5.81 mg/mL
Injected sample volume: 50 uL	**Mass Recovered (By RI):** 214 ug	**% Recovery (By RI):** 73.67 %
Recoverable Stock Conc.: 4.28 ug/uL	**Monodispersity Index:** 0 %	
Notes:		

FIG. 8. Aggregation screening. A combination of analytical gel filtration, static light scattering, and refractive index detects the volume and mass of each protein species in solution. (A) The static light scattering of NESG target protein BlR15 in a ''no salt'' buffer is shown in the top chromatogram and indicates a single peak under these buffer conditions. The bottom chromatogram traces the refractive index; the single peak in the lower molecular weight region (corresponding to the single peak in the light scattering) shows that most of the mass injected into the column is contained in this peak, indicating that the majority of the protein in this buffer is monomeric. (B) The Aggregation Screening results for NESG target BlR15 are summarized in this view from SPINE.

additives) can be varied in a high-throughput manner and tested for their ability to produce an environment favorable for a monodisperse population. More specifically, IMAC-purified protein is exchanged into a series of several different buffers through an overnight dialysis at 4°. Each sample containing the protein exchanged into a specific buffer is then injected into the FPLC, and the species present in solution are separated based on size using a Shodex gel filtration column. The radius and mass of the species responsible for each peak are computed in real time. The aggregation state of the protein is thus characterized under different buffer conditions, and the buffer promoting the highest degree of monodispersity is chosen for the next step of preparative gel filtration purification.

Gel Filtration Chromatography

Preparative-scale gel filtration chromatography is generally performed under the buffer conditions that most favor monodispersity. The Rutgers Protein Production facility has a series of AKTA FPLC (Amersham Biosciences) chromatography systems including four Primes, one Purifier, one Explorer, and finally a four module AKTAxpress system. Each system is configured with a HiLoad 26/60 Superdex 75 gel filtration column(s) (Amersham Biosciences), with the capacity to safely load up to 10 ml of sample volume. Gel filtration columns are first equilibrated with buffers generally favoring monodispersity; for example, no salt (NS) buffer (10 m*M* Tris, 5 m*M* DTT, pH 7.5) promotes monodispersity for target BlR15 (Fig. 8). Protein samples are loaded onto the column, and the peak containing the monodisperse protein is collected. Gel filtration purified samples are then either prepared for NMR data collection or concentrated to ~10 mg/ml using an Ultrafree Centrifugal Filter Unit (Millipore), flash frozen in small (50-μl) aliquots to minimize the effects of protein dehydration upon freezing, and stored at −80°. These frozen samples are shipped to collaborators for crystallization screening.

"Button Tests" to Optimize Conditions for NMR Studies

One striking statistic from the HSQC screening is the fact that more than 25% of the samples produced have "good" (i.e., promising, good, or excellent) HSQC scores, indicating their structure is likely solvable by NMR. However, many of the targets that initially produce good spectra exhibit various forms of sample instability during data collection, including proteolysis, oxidation, deamidation, and slow precipitation between the time they are prepared and the completion of NMR data collection. Different temperatures and buffer conditions can produce significant

differences in both spectral quality and sample stability. For example, in some cases cocktails of protease inhibitors are added to inhibit proteolysis during NMR data collection. Generally, even for a ''high-throughput screening'' platform, it is necessary to optimize buffer conditions before committing to significant NMR data collection time.

Currently, the sample stability issue that is most limiting to NMR structure production is the *slow precipitation* of subject proteins in NMR samples. Some 25% of our gel filtration–purified NMR samples exhibit slow precipitation over days or weeks after the sample is prepared, which can severely frustrate data collection efforts. To screen for conditions that avoid this *slow precipitation* behavior, we have implemented microscale buffer screening using microdialysis buttons (Bagby *et al.*, 1997) to identify conditions that stabilize the protein preparations. One advantage of this system is the small amount of protein sample needed for analysis (~50 μg for a dozen conditions), allowing a large range of conditions to be tested, including pH, salt, and other additives at varying concentrations. Table III lists 12 conditions, corresponding to three pH values, and the presence or absence of NaCl, L-arginine, and DTT. An individual microdialysis button of the protein sample is dialyzed against each of these buffers at 4°, and these ''buttons'' are then examined for signs of protein precipitation overnight using a dissecting microscope, and again after 1 week; the button is then moved to the same buffer at 20° and observed for another 2 weeks. We have often found that 100 m*M* L-arginine is useful for stabilizing protein samples against slow precipitation. Buffer conditions promoting sample stability identified from this assay are combined with information

TABLE III
BUFFER CONDITIONS OF INITIAL TESTS FOR STABILITY WITH RESPECT TO SLOW PROTEIN PRECIPITATION

Buffer	NaCl	DTT	Arginine
50 m*M* ammonium acetate, pH 5.0	0	0	0
50 m*M* ammonium acetate, pH 5.0	0	10 m*M*	0
50 m*M* ammonium acetate, pH 5.0	0.1 *M*	10 m*M*	0
50 m*M* ammonium acetate, pH 5.0	0	10 m*M*	0.1 *M*
50 m*M* MES, pH 6.0	0	0	0
50 m*M* MES, pH 6.0	0	10 m*M*	0
50 m*M* MES, pH 6.0	0.1 *M*	10 m*M*	0
50 m*M* MES, pH 6.0	0	10 m*M*	0.1 *M*
50 m*M* Bis. Tris, pH 6.5	0	0	0
50 m*M* Bis. Tris, pH 6.5	0	10 m*M*	0
50 m*M* Bis. Tris, pH 6.5	0.1 *M*	10 m*M*	0
50 m*M* Bix. Tris, pH 6.5	0	10 m*M*	0.1 *M*

from the NMR screening data to ascertain optimal buffer conditions for NMR data collection.

Quality Control

Proteins prepared for NMR or X-ray crystallographic studies are all analyzed for homogeneity by SDS–PAGE and validated for molecular weight by matrix-assisted laser-desorption-induced time-of-flight (MALDI-TOF) mass spectrometry. When inconsistencies are observed, the expression constructs are validated by DNA sequencing. Data generated at each stage of the production pipeline, along with analytical results, spectra, comments, records of interlaboratory shipments, the names of the individuals involved in each production step, and other aspects of the production process, are archived in the SPINE database. Summaries of these data are available in public domain (http://nesg.org/).

Capacity of the Platform

Based on our current levels of success in producing diffraction quality crystals or samples amenable for NMR studies, we calculate that to reach our goal of determining 100–200 novel structures per year requires the capacity to produce 600 target proteins on the 10–50 mg scale per year, each with high (>98%) homogeneity. Our conservative estimate with current protein target list characteristics suggests that this requires producing roughly 2500 expression constructs per year. In the sections above, we outlined a scalable platform for high-throughput protein production of samples suitable for structural determination, including the needed technologies and infrastructure. The platform as it stands is producing these target numbers of expression constructs (2500 per year) and purified proteins (600 per year in 10–50 mg quantities) for both NMR and X-ray crystallization experiments.

Hierarchical Multiplex Expression

The ability to clone and analyze the expression of targets in a high-throughput manner has also allowed us to become more proficient and successful at rescuing targets that were not initially suitable for structural determination because of low expression, low solubility, poor NMR spectral quality, or poor quality crystals. The first layer of our hierarchical multiplex strategy is the use of multiple homologues from a particular target family (Fig. 1). This process is managed through the SPINE database. Next, the platform allows for cloning targets into a series of expression vectors with different placements of the affinity tag or promoters driving expression.

For example, a 96-well plate was recently assembled with targets ranging from bacteria to human proteins, all of which were previously cloned into a C-terminal 6X-His expression vector and were either not expressed or insoluble. These targets were then rapidly subcloned into an N-terminal expression vector (using a single universal primer and our robotic platform). Over 30% of the resulting N-terminal 6XHis-tagged proteins were expressed and soluble at levels amenable to preparative-scale expression and purification.

Several other new technologies are being explored for expanding our hierarchical-multiplex expression platform. *E. coli*–based cold-shock induction vectors (Qing *et al*., 2004) have allowed production of many targets that are not expressed or insoluble in pET-derived expression vectors. High-throughput robotic-based expression technology also allows for varying other parameters of the expression conditions, such as the bacterial host strains, and efforts are in progress to explore and develop improved chaperone-supplemented bacterial host strains. Efforts are also in progress to develop robotic technologies for protein production in cell-free wheat germ (Ma *et al*., 2005; Morita *et al*., 2003; Sawasaki *et al*., 2002), *Pichia pastoris* (Boettner *et al*., 2002; Prinz *et al*., 2004; Wood and Komives, 1999), and *Saccharomyces cervisiae* (Boettner *et al*., 2002; Holz *et al*., 2003; Prinz *et al*., 2004) expression systems, particularly for eukaryotic protein targets. Accordingly, by hierarchical multiplexing of expression technologies, a significant number of targets that have not passed current analyses may eventually be produced in a form amenable to 3D structure determination.

Data Integration and Sharing

The SPINE (Bertone *et al*., 2001; Goh *et al*., 2003) and SPINS (Baran *et al*., 2002) databases collect information from all steps in the protein production process, including information on the cloning and small-scale expression, large-scale fermentation and protein preparations, aggregation screening, and NMR screening. SPINE also tracks sample shipments between laboratories of the NESG Consortium. It is a central component of our protein production pipeline, acting to integrate the entire process of sample production and analysis and to organize data that will be invaluable for optimizing the sample production process and learning about physical and biochemical properties of proteins.

Summary

We have outlined our strategy and platform for producing high-quality protein samples using *E. coli* expression hosts. Our protein purification process is centered on 6X-His affinity tag-IMAC affinity purification,

allowing all of our targets to have identical initial purification procedures and the implementation of high-throughput parallel methods. In most cases, these 6X-His-tagged proteins are sufficiently pure that a single ensuing gel filtration chromatography step is adequate to produce protein preparations that are greater than 98% homogeneous, a level that we have observed is sufficient for structural studies. Protein structures have generally been determined by NMR and X-ray crystallography with the small 6X-His tags on the protein targets, although in a few cases the tags have been removed by cloning into related vectors that provide tagless proteins. Our targets include primarily proteins that comprise the proteomes of the eukaryotic model organisms and their prokaryotic homologues (Fig. 9), and our current strategies have proven effective in producing samples containing structured eukaryotic and prokaryotic proteins.

The target list of the NESG Consortium (Liu and Rost, 2004; Liu *et al.*, 2004; Wunderlich *et al.*, 2004), roughly two-thirds of which are eukaryotic proteins (Fig. 9), is generally more challenging than those pursued by structural genomics projects focused exclusively on prokaryotic proteins, which tend to be easier to produce in bacterial host systems. Despite these challenges, over 1000 different protein targets have been cloned, expressed, and purified in tens of milligram quantities over the past 36-month period (see Summary Statistics for All Targets, http://www-nmr.cabm.rutgers.edu/bioinformatics/ZebaView/) in the Rutgers facility; current production rates

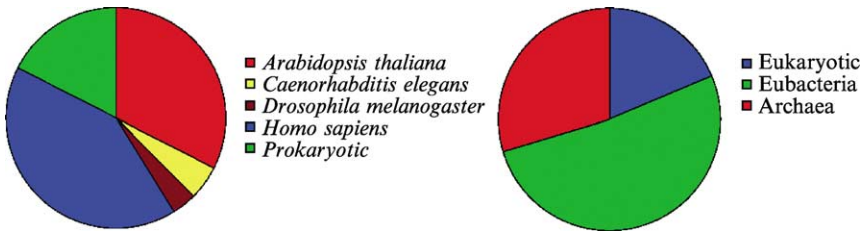


FIG. 9. (*Left*) Phylogenetic distribution of NESG target proteins. Currently, 80% of the NESG targets are from the eukaryotic model organisms *Arabidopsis thaliana, Caenorhabditis elegans, Drosophila melanogaster*, and *Homo sapiens*. As illustrated in the pie chart, the majority of proteins are derived from the human and *Arabidopsis* genomes. The remaining 20% of NESG targets for Rutgers protein production (*S. cerevisiae* efforts are focused in Toronto) are of prokaryotic origin, including proteins from both archaea and eubacteria; see Table I for a complete listing of the eukaryotic organisms and the number of proteins targeted from each of these proteomes. (*Right*) Phylogenetic distribution of NESG protein structures deposited in the PDB. As indicated in the pie chart, ~20% of the NESG structures are of eukaryotic proteins, ~30% are of archaeal origin, and the remaining are structures of proteins from eubacteria. However, the majority of these prokaryotic proteins for which structures have been determined are members of protein domain families that also include eukaryotic members.

are about 12 purified protein targets in tens of milligram quantities per week. Our experience using the hierarchical multiplex expression and purification strategy, also described in this chapter, has allowed us to achieve success in producing not only protein samples but also three-dimensional structures. As of December 2004, the NESG Consortium has deposited over 145 new protein structures to the PDB; about two-thirds of these protein samples were produced by the Rutgers NESG Protein Production Facility. Roughly 20% of the NESG protein structures are of eukaryotic proteins (Fig. 9), demonstrating the broad applicability of the platform. The sample production and screening technologies described here are scalable and, as demonstrated by several other chapters in this volume, efficiencies of sample production and structure determination methods in use by the NESG Consortium continue to improve. In addition to their role in our pilot structural genomics initiative, these improved robotic and/or parallel cloning, expression, protein production, and biophysical screening technologies will be of broad value to the structural biology, functional proteomics, and structural genomics communities.

Acknowledgments

We thank Drs. S. Anderson, P. Manor, F. Piano, and A. Yee for helpful advice in developing this protein production platform. This work is supported by Grant P50-GM62413 from the Protein Structure Initiative of the National Institutes of Health, Institute of General Medical Sciences.

References

Abremski, K., and Hoess, R. (1984). Bacteriophage P1 site-specific recombination. Purification and properties of the Cre recombinase protein. *J. Biol. Chem.* **259,** 1509–1514.

Acton, T. B., Gunsalus, K., Xiao, R., Ma, L., Chiang, Y., Clement, T., Everett, J. K., Shastry, R., Denissova, N., Palacios, D., *et al.* (2005). The protein sample production platform of the Northeast Structural Genomics Consortium. *J. Struct. Funct. Genomics.* Submitted.

Bagby, S., Tong, K. I., Liu, D., Alattia, J. R., and Ikura, M. (1997). The button test: A small scale method using microdialysis cells for assessing protein solubility at concentrations suitable for NMR. *J. Biomol. NMR* **10,** 279–282.

Baran, M. C., Moseley, H. N., Sahota, G., and Montelione, G. T. (2002). SPINS: Standardized protein NMR storage. A data dictionary and object-oriented relational database for archiving protein NMR spectra. *J. Biomol. NMR* **24,** 113–121.

Baran, M. C., Haung, Y. J., Moseley, H. N., and Montelione, G. T. (2004). Automated analysis of protein NMR assignments and structures. *Chem. Rev.* **104,** 3541–3556.

Bertone, P., Kluger, Y., Lan, N., Zheng, D., Christendat, D., Yee, A., Edwards, A. M., Arrowsmith, C. H., Montelione, G. T., and Gerstein, M. (2001). SPINE: An integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics. *Nucleic Acids Res.* **29,** 2884–2898.

Boettner, M., Prinz, B., Holz, C., Stahl, U., and Lang, C. (2002). High-throughput screening for expression of heterologous proteins in the yeast *Pichia pastoris*. *J. Biotechnol.* **99,** 51–62.

Bradford, M. M. (1976). A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein dye binding. *Anal. Biochem.* **131,** 248–254.

Bujard, H., Gentz, R., Lanzer, M., Stuber, D., Muller, M., Ibrahimi, I., Hauptle, M. T., and Dobberstein, B. (1987). A T5 promotor based transcription-translation system for the analysis of proteins *in vivo* and *in vitro*. *Methods Enzymol.* **155,** 416–433.

Carter, P., Liu, J., and Rost, B. (2003). PEP: Predictions for entire proteomes. *Nucleic Acids Res.* **31,** 410–413.

Chen, G. F., and Inouye, M. (1990). Suppression of the negative effect of minor arginine codons on gene expression; preferential usage of minor codons within the first 25 codons of the *Escherichia coli* genes. *Nucleic Acids Res.* **18,** 1465–1473.

Crowe, J., Dobeli, H., Gentz, R., Hochuli, E., Stuber, D., and Henco, K. (1994). 6xHis-Ni-NTA chromatography as a superior technique in recombinant protein expression/purification. *Methods Mol. Biol.* **31,** 371–387.

Dodson, E. (2003). Is it jolly SAD? *Acta Crystallogr D Biol. Crystallogr.* **59,** 1958–1965.

Doublie, S., Kapp, U., Aberg, A., Brown, K., Strub, K., and Cusack, S. (1996). Crystallization and preliminary X-ray analysis of the 9 kDa protein of the mouse signal recognition particle and the selenomethionyl-SRP9. *FEBS Lett.* **384,** 219–221.

Everett, J. K., Acton, T. B., and Montelione, G. T. (2004). Primer Prim'r: A web based server for automated primer design. *J. Struct. Funct. Genomics* **5,** 13–21.

Ferre-D'Amare, A. R., and Burley, S. K. (1994). Use of dynamic light scattering to assess crystallizability of macromolecules and macromolecular assemblies. *Structure* **2,** 357–359.

Ferre-D'Amare, A. R., and Burley, S. K. (1997). Dynamic light scattering in evaluating crystallizability of macromolecule. *Methods Enzymol.* **276,** 157–166.

Geisse, S., Gram, H., Kleuser, B., and Kocher, H. P. (1996). Eukaryotic expression systems: A comparison. *Protein Expr. Purif.* **8,** 271–282.

Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y. L., Ooi, C. E., Godwin, B., Vitols, E., *et al.* (2003). A protein interaction map of *Drosophila melanogaster*. *Science* **302,** 1727–1736.

Goh, C. S., Lan, N., Echols, N., Douglas, S. M., Milburn, D., Bertone, P., Xiao, R., Ma, L. C., Zheng, D., Wunderlich, Z., *et al.* (2003). SPINE 2: A system for collaborative structural proteomics within a federated database framework. *Nucleic Acids Res.* **31,** 2833–2838.

Hartley, J. L., Temple, G. F., and Brasch, M. A. (2000). DNA cloning using *in vitro* site-specific recombination. *Genome Res.* **10,** 1788–1795.

Hendrickson, W. A., and Ogata, C. M. (1997). Phase determination from multiwavelength anomalous diffraction measurements. *Methods Enzymol.* **276,** 494–523.

Hendrickson, W. A., Horton, J. R., and LeMaster, D. M. (1990). Selenomethionyl proteins produced for analysis by multiwavelength anomalous diffraction (MAD): A vehicle for direct determination of three-dimensional structure. *EMBO J.* **9,** 1665–1672.

Holz, C., Prinz, B., Bolotina, N., Sievert, V., Bussow, K., Simon, B., Stahl, U., and Lang, C. (2003). Establishing the yeast *Saccharomyces cerevisiae* as a system for expression of human proteins on a proteome-scale. *J. Struct. Funct. Genomics* **4,** 97–108.

Ikemura, T. (1985). Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.* **2,** 13–34.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA* **98,** 4569–4574.

Jansson, M., Li, Y.-C., Jendenberg, L., Anderson, S., and Montelione, G. T. (1996). High-level production of uniformly ^{15}N- and ^{13}C-enriched fusion proteins in *Escherichia coli*. *J. Biomol. NMR* **7,** 131–141.

Kapust, R. B., Tozser, J., Copeland, T. D., and Waugh, D. S. (2002). The P1′ specificity of tobacco etch virus protease. *Biochem. Biophys. Res. Commun.* **294,** 949–955.

Lander, E. S. (1999). Array of hope. *Nat. Genet.* **21,** 3–4.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921.

Li, S., Armstrong, C. M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P. O., Han, J. D., Chesneau, A., Hao, T., *et al.* (2004). A map of the interactome network of the metazoan *C. elegans*. *Science* **303,** 540–543.

Liu, J., and Rost, B. (2004). CHOP proteins into structural domain-like fragments. *Proteins* **55,** 678–688.

Liu, J., Hegyi, H., Acton, T. B., Montelione, G. T., and Rost, B. (2004). Automatic target selection for structural genomics on eukaryotes. *Proteins* **56,** 188–200.

Ma, L. C., Sawasaki, T., Tsuchimochi, M., Mazda, S., Gunsalus, K. C., Macapagal, D., Shastry, R., Ho, C. K., Acton, T. B., Endo, Y., and Montelione, G. T. (2005). Evaluation of a wheat germ cell-free protein production system for expression and solubility screening of eukaryotic proteins. *J. Struct. Funct. Genomics.* Submitted.

Makrides, S. C. (1996). Strategies for achieving high-level expression of genes in *Escherichia coli*. *Microbiol. Rev.* **60,** 512–538.

Manor, P., Shen, J., Satterwhite, R., Kuzin, A., Forohar, F., Benach, J., Smith, P., Montelione, G. T., Acton, T. B., and Hunt, J. (2005). *Protein solution aggregation characteristics and crystallization.* In preparation.

Millard, C. S., Stols, L., Quartey, P., Kim, Y., Dementieva, I., and Donnelly, M. I. (2003). A less laborious approach to the high-throughput production of recombinant proteins in *Escherichia coli* using 2-liter plastic bottles. *Protein Expr. Purif.* **29,** 311–320.

Montelione, G. T., and Anderson, S. (1999). Structural genomics: Keystone for a Human Proteome Project. *Nat. Struct. Biol.* **6,** 11–12.

Morita, E. H., Sawasaki, T., Tanaka, R., Endo, Y., and Kohno, T. (2003). A wheat germ cell-free system is a novel way to screen protein folding and function. *Protein Sci.* **12,** 1216–1221.

Prinz, B., Schultchen, J., Rydzewski, R., Holz, C., Boettner, M., Stahl, U., and Lang, C. (2004). Establishing a versatile fermentation and purification procedure for human proteins expressed in the yeasts *Saccharomyces cerevisiae* and *Pichia pastoris* for structural genomics. *J. Struct. Funct. Genomics* **5,** 29–44.

Qing, G., Ma, L., Khorchid, A., Swapna, G. V. T., Mal, T. K., Takayama, M. M., Xia, B., Sangita Phadtare, S., Ke, H., Acton, T., *et al.* (2004). Cold-shock induced high-yield protein production in *Escherichia coli*. *Nat. Biotechnol.* **22,** 877–882.

Rain, J. C., Selig, L., De Reuse, H., Battaglia, V., Reverdy, C., Simon, S., Lenzen, G., Petel, F., Wojcik, J., Schachter, V., *et al.* (2001). The protein-protein interaction map of *Helicobacter pylori*. *Nature* **409,** 211–215.

Reboul, J., Vaglio, P., Rual, J. F., Lamesch, P., Martinez, M., Armstrong, C. M., Li, S., Jacotot, L., Bertin, N., Janky, R., *et al.* (2003). *C. elegans* ORFeome version 1.1: Experimental verification of the genome annotation and resource for proteome-scale protein expression. *Nat. Genet.* **34,** 35–41.

Rubin, G. M., Hong, L., Brokstein, P., Evans-Holm, M., Frise, E., Stapleton, M., and Harvey, D. A. (2000). A *Drosophila* complementary DNA resource. *Science* **287,** 2222–2224.

Sauer, B. (1994). Site-specific recombination: Developments and applications. *Curr. Opin. Biotechnol.* **5,** 521–527.

Sawasaki, T., Ogasawara, T., Morishita, R., and Endo, Y. (2002). A cell-free protein synthesis system for high-throughput proteomics. *Proc. Natl. Acad. Sci. USA* **99,** 14652–14657.

Sheibani, N. (1999). Prokaryotic gene fusion expression systems and their use in structural and functional studies of proteins. *Prep. Biochem. Biotechnol.* **29,** 77–90.

Shirano, Y., and Shibata, D. (1990). Low temperature cultivation of *Escherichia coli* carrying a rice lipoxygenase L-2 cDNA produces a soluble and active enzyme at a high level. *FEBS Lett.* **271,** 128–130.

Shuman, S. (1994). Novel approach to molecular cloning and polynucleotide synthesis using vaccinia DNA topoisomerase. *J. Biol. Chem.* **269,** 32678–32684.

Sorensen, M. A., Kurland, C. G., and Pedersen, S. (1989). Codon usage determines translation rate in *Escherichia coli*. *J. Mol. Biol.* **207,** 365–377.

Stapleton, M., Carlson, J., Brokstein, P., Yu, C., Champe, M., George, R., Guarin, H., Kronmiller, B., Pacleb, J., Park, S., *et al.* (2002). A *Drosophila* full-length cDNA resource. *Genome Biol.* **3,** 80–88.

Strausberg, R. L., Feingold, E. A., Klausner, R. D., and Collins, F. S. (1999). The mammalian gene collection. *Science* **286,** 455–457.

Strausberg, R. L., Feingold, E. A., Grouse, L. H., Derge, J. G., Klausner, R. D., Collins, F. S., Wagner, L., Shenmen, C. M., Schuler, G. D., Altschul, S. F., *et al.* (2002). Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc. Natl. Acad. Sci. USA* **99,** 16899–16903.

Studier, F. W. (2004). Personal communication.

Studier, F. W., Rosenberg, A. H., Dunn, J. J., and Dubendorff, J. W. (1990). Use of T7 RNA polymerase to direct expression of cloned genes. *Methods Enzymol.* **185,** 60–89.

Tartof, K. D., and Hobbs, C. A. (1987). Improved media for growing plasmid and cosmid clones. *Bethesda Res. Lab. Focus* **9,** 12.

Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., *et al.* (2000). A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403,** 623–627.

Winzeler, E. A., Schena, M., and Davis, R. W. (1999). Fluorescence-based expression monitoring using microarrays. *Methods Enzymol.* **306,** 3–18.

Wood, M. J., and Komives, E. A. (1999). Production of large quantities of isotopically labeled protein in *Pichia pastoris* by fermentation. *J. Biomol. NMR* **13,** 149–159.

Wunderlich, Z., Acton, T. B., Liu, J., Kornhaber, G., Everett, J., Carter, P., Lan, N., Echols, N., Gerstein, M., Rost, B., and Montelione, G. T. (2004). The protein target list of the Northeast Structural Genomics Consortium. *Proteins* **56,** 181–187.

Yee, A., Chang, X., Pineda-Lucena, A., Wu, B., Semesi, A., Le, B., Ramelot, T., Lee, G. M., Bhattacharyya, S., Gutierrez, P., *et al.* (2002). An NMR approach to structural proteomics. *Proc. Natl. Acad. Sci. USA* **99,** 1825–1830.

Yee, A., Pardee, K., Christendat, D., Savchenko, A., Edwards, A. M., and Arrowsmith, C. H. (2003). Structural proteomics: Toward high-throughput structural biology as a tool in functional genomics. *Acc. Chem. Res.* **36,** 183–189.