Published in final edited form as:

Methods Enzymol. 2011 ; 493: 21–60. doi:10.1016/B978-0-12-381274-2.00002-9.

Preparation of Protein Samples for NMR Structure, Function, and Small Molecule Screening Studies

Thomas B. Acton¹, Rong Xiao¹, Stephen Anderson¹, James Aramini¹, William A. Buchwald¹, Colleen Ciccocanti¹, Ken Conover¹, John Everett¹, Keith Hamilton¹, Yuanpeng Janet Huang¹, Haleema Janjua¹, Gregory Kornhaber¹, Jessica Lau¹, Dong Yup Lee¹, Gaohua Liu¹, Melissa Maglaqui¹, Lichung Ma¹, Lei Mao¹, Dayaban Patel¹, Paolo Rossi¹, Seema Sahdev¹, Ritu Shastri², G.V.T. Swapna¹, Yeufeng Tang¹, Saichiu Tong¹, Dongyan Wang¹, Huang Wang¹, Li Zhao¹, and Gaetano T. Montelione^{1,2,*}

¹Center for Advanced Biotechnology and Medicine, Department of Molecular Biology and Biochemistry, and Northeast Structural Genomics Consortium, Rutgers University, Piscataway, New Jersey 08854

²Department of Biochemistry, Robert Wood Johnson Medical School, University of Medicine and Dentistry of New Jersey, Piscataway, NJ 08854

Abstract

In this chapter, we concentrate on the production of high quality protein samples for NMR studies. In particular, we provide an in-depth description of recent advances in the production of NMR samples and their synergistic use with recent advancements in NMR hardware. We describe the protein production platform of the Northeast Structural Genomics Consortium, and outline our high-throughput strategies for producing high quality protein samples for nuclear magnetic resonance (NMR) studies. Our strategy is based on the cloning, expression and purification of 6X-His-tagged proteins using T7-based *Escherichia coli* systems and isotope enrichment in minimal media. We describe 96-well ligation-independent cloning and analytical expression systems, parallel preparative scale fermentation, and high-throughput purification protocols. The 6X-His affinity tag allows for a similar two-step purification procedure implemented in a parallel high-throughput fashion that routinely results in purity levels sufficient for NMR studies (> 97% homogeneity). Using this platform, the protein open reading frames of over 17,500 different targeted proteins (or domains) have been cloned as over 28,000 constructs. Nearly 5,000 of these proteins have been purified to homogeneity in tens of milligram quantities (see Summary Statistics, <http://nesg.org/statistics.html>), resulting in more than 950 new protein structures, including more than 400 NMR structures, deposited in the Protein Data Bank. The Northeast Structural Genomics Consortium pipeline has been effective in producing protein samples of both prokaryotic and eukaryotic origin. Although this paper describes our entire pipeline for producing isotope-enriched protein samples, it focuses on the major updates introduced during the last 5 years (Phase 2 of the National Institute of General Medical Sciences Protein Structure Initiative). Our advanced automated and/or parallel cloning, expression, purification, and biophysical screening technologies are suitable for implementation in a large individual laboratory or by a

*Correspondence should be addressed to: Prof. Gaetano T. Montelione, CABM-Rutgers University, 679 Hoes Lane, Piscataway, NJ 08854, Ph. 732-235-5321; guy@cabm.rutgers.edu.

small group of collaborating investigators for structural biology, functional proteomics, ligand screening and structural genomics research.

Keywords

Structural Genomics; High throughput protein production; Construct optimization; Disorder prediction; Ligation independent cloning; Multiple Displacement Amplification; Laboratory Information Management System; Protein Structure Initiative; NMR; T7 Escherichia coli expression system; Wheat Germ Cell Free; NMR microprobe screening; Parallel protein purification; 6X-His tag; HDX-MS; Total gene synthesis; condensed single protein production

Introduction

The production of high quality protein samples is critical to success in structural biology and drug discovery. During the second phase of the Protein Structure Initiative, the Northeast Structural Genomics Consortium (NESG; <http://www.nesg.org>) was one of four Large Scale Centers funded by the National Institutes of Health, National Institute of General Medical Sciences. The goal of these centers was to determine the three-dimensional atomic-level structures of hundreds of novel proteins and protein domains and leverage this novel structural information to allow three-dimensional modeling of thousands of additional proteins (or protein domains). Another major goal of these centers was to develop and refine new technologies for high-throughput protein production, X-ray crystallography, NMR spectroscopy, structural bioinformatics, and related supporting infrastructure. The Protein Structure Initiative has worked to advance the field of biology through the dissemination of three-dimensional structural information on important protein domain families, production of protein expression systems and protocols, and by providing improved technology for protein sample preparation and structural analysis.

The protein sample production pipeline of the NESG has been previously described in detail (Acton et al., 2005; Xiao et al., 2010). Unlike nucleic acid based genomics studies, where the macromolecules share common biophysical traits and standardized preparation techniques, proteins have a wide range of biophysical properties making high-throughput production and sample purification significantly more challenging. Adding further complexity is the need to produce selenomethionine labeled proteins for X-ray crystallography studies and isotope-enriched samples for nuclear magnetic resonance (NMR) studies. One of the unique features of the NESG pipeline is the ability to produce protein samples suitable for both structural determination strategies. Indeed the NESG has produced a similar number of three-dimensional structures by each method. However, in this chapter we concentrate on the production of high quality protein samples for NMR studies. In particular, we provide an in-depth description of recent advances in the production of NMR samples and their synergistic use with recent advancements in NMR hardware. These include NMR microprobe technology allowing determination of protein structures from 100 microgram quantities of protein (Aramini et al., 2007), microprobe screening of small quantities of proteins in order to assess amenability to NMR structural determination, buffer optimization for protein stability and/or spectral quality (Rossi et al., 2010), wheat germ cell-free protein translation for the production of protein samples (Vinarov et al., 2006; Zhao

et al., 2010), and condensed-phase expression technologies allowing reduced costs in preparing isotope-enriched protein samples (Schneider et al., 2009, 2010).

The NESG high-throughput cloning, protein expression and protein purification pipeline is primarily based on *E. coli* T7 expression systems (Studier and Moffatt, 1986), utilizing 6X-His tags to allow a similar purification strategy to be utilized for proteins with diverse biophysical characteristics. Overall this approach has proven to be a highly productive, efficient, and inexpensive method to produce the quantities of protein required for structural studies (Graslund et al., 2008a). We also describe our strategies for target selection, construct optimization, ligation-independent cloning, analytical scale expression and solubility screening, midi-scale expression, purification and biophysical characterization and large-scale protein sample production. The protein targets of the NESG project are both full-length proteins and domain constructs, we describe experimental and prediction methods for identifying disordered regions of proteins, and how this information is used to design protein constructs suitable for NMR studies.

The current output of the NESG protein production facility on a weekly basis includes cloning and expression screening of over one hundred protein targets, fermenting on a preparative (1 – 2 L) scale 50-75 expression constructs, and purifying in tens of milligram quantities roughly 30 - 40 targets for biophysical characterization, including NMR and/or crystallization screening. This platform can be readily implemented by traditional structural biology laboratories, biotechnology industry, and various proteomics and functional genomics projects since it is scaleable, portable, and largely comprised of commercially available equipment.

1. Bioinformatics Infrastructure and Target Curation

During the second phase of the Protein Structure Initiative program, a centralized bioinformatics committee selected and distributed protein targets to the Large Scale Centers (Dessailly et al., 2009). These generally constituted broadly conserved protein domain families for which no structural representative was yet available (BIG families), very large protein domain families with limited structural coverage (MEGA families), and domain families selected from metagenomic projects (META-families) such as the human gut microbiome project (Gill et al., 2006). The overall goal of targeting large protein domain families is to provide the greatest novel leverage of structure space per target (Liu et al., 2007; Nair et al., 2009). Fortuitously, this allows for the highly effective pan-genomic targeting strategy, taking advantage of the sequence differences and their concomitant biophysical characteristics within a domain family to identify and pursue the members most amenable for structure determination (Liu et al., 2004; Acton et al., 2005; Punta et al., 2009). Each Large Scale Center also defined a biomedical theme; e.g. the NESG is pursuing structural analysis of large numbers of proteins from the Human Cancer Pathway Protein Interaction Network (Huang et al., 2008). These proteins are involved in cancer associated signaling pathways and biological processes, together with their associated protein-protein interaction partners (<http://nesg.org:9090/HCPIN/>). In addition, the biomedical community nominates targets to the Protein Structure Initiative (http://www.nesg.org/target_nomination.html), which distributes these “Community Nominated Targets” to the

various Large Scale Centers. Irrespective of the varied source of protein targets, the focus of the NESG remains the domain families represented in eukaryotic proteomes, including families that have exclusively eukaryotic members (e.g. the Ubiquitin Domain family) and families that have both eukaryotic and prokaryotic members (e.g. the START Domain family).

One of the main tasks of the Protein Structure Initiative and structural genomics as a whole is to increase the efficiency of protein three-dimensional structure production, ranging from target selection to automated Protein Data Bank depositions. With respect to protein production, experimental and bioinformatics studies have been undertaken to identify the parameters and procedures that correlate with success, such as high levels of protein solubility or “clone to Protein Data Bank deposition rates” (Goh et al., 2004; Slabinski et al., 2007; Graslund et al., 2008b). The NESG has developed numerous bioinformatics tools to take advantage of pan-genomic targeting and identify the members of a protein domain family that are most amenable to protein production and structure determination. It is now well established that variation in protein sequence within a protein domain family can greatly affect its biophysical properties, and therefore its success in protein production. The NESG has collected a vast data set on the behavior of proteins from diverse sources and families (both eukaryotic and prokaryotic) with the important fact that they were all prepared in a similar fashion. This has allowed NESG researchers to identify primary sequence traits that correlate with (i) high levels of protein expression (P_{ES}) (unpublished results), (ii) greater probability of crystal structure determination based on protein sequence (P_{XS}) (Price et al., 2009) (<http://nmr.cabm.rutgers.edu:8080/PXS/calculatePXS.jsp>) and (iii) greater probability of amenability to NMR structure determination (P_{NMR}) (unpublished results). Using our extensive list of over 175 Reagent Genomes (fully sequenced archeal, bacterial, and eukaryotic genomes and the corresponding genetic material for cloning) and these tools, we identify members of a protein family that are most likely to proceed to structure determination. This allows us to select several (4 – 6) proteins from each family for protein production in an effort to take advantage of the pan genomic targeting strategy without pursuing large numbers of targets from a given protein family.

Beyond increasing efficiency by enriching our protein production pipeline with amenable targets, another major enhancement to our pipeline in Protein Structure Initiative-2 is the NESG Construct Optimization Software. Highly homogeneous protein samples with minimal numbers of disordered residues are generally more amenable for successful protein crystallization and structure determination by X-ray crystallography (Sharma et al., 2009). Although this chapter is focused on sample preparation for NMR studies, which can often be used successfully to study even fully disordered proteins, disordered segments of proteins can promote aggregation and deleteriously affect NMR spectral quality. In addition, a large percentage of targets (particularly human and other eukaryotic targets) are within multidomain proteins, which often misfold in prokaryotic systems (Netzer and Hartl, 1997). Of even more importance, many multidomain proteins exceed the size limitations for high-throughput NMR structural determination techniques. Domain parsing can be used to circumvent these significant issues. However it is extremely challenging to predict the protein subsequence that will produce a soluble well-behaved protein, particularly with

domains for which the three-dimensional structure is not yet known. This arises from problems with accurately predicting the domain boundaries and locations of disordered residues, and how to use such information to design an open reading frame that produces expression and solubility in the T7 system. Currently our approach is to take advantage of our high-throughput platform and produce several alternative constructs, varying the termini of a targeted domain, followed by experiments to identify the protein subsequence with the best behavior. The success of this strategy has been reported by the NESG and others (Graslund et al., 2008b; Chikayama et al., 2010; Xiao et al., 2010).

The NESG construct optimization software uses reports from the DisMeta Server (<http://www-nmr.cabm.rutgers.edu/bioinformatics/disorder/>), a metaserver that generates a consensus analysis of eight sequence-based disorder predictors to identify regions that are likely to be disordered. It also identifies predicted secretion signal peptides using SignalP (Bendtsen et al., 2004), trans-membrane segments by TMHMM (Krogh et al., 2001), possible metal binding sites (Bertini et al., 2010), secondary structure by PROFsec (Rost et al., 2004) and PSIPred (McGuffin et al., 2000), and interdomain disordered linkers (Figure 1A). The data from these prediction servers, along with multiple sequence alignments of homologous proteins and hidden Markov models characteristic of the targeted protein domain families (Dessailly et al., 2009), are used to predict possible structural domain boundaries. Based on this information, the software generates nested sets of alternative constructs, for full-length proteins, multidomain constructs, and single domain constructs. Thus for a single targeted region, we generally design multiple open reading frames varying the N and/or C-terminal sequences (Figure 1B). These alternative constructs often possess significantly better expression, solubility and biophysical behavior than their full-length parent sequences, increasing the likelihood of success in crystallization and the efficiency of structure production. In addition, it generates domain-sized regions that are amenable to high-throughput NMR studies, allowing access to proteins that would otherwise be too large to study by NMR.

An example of our domain parsing / construct optimization approach is shown Figure 1A. Consensus analysis of several disorder prediction algorithms (see Disorder Consensus panel) suggests that the C-terminal half of the 434-residue protein from *Porphyromonas gingivalis* Q7MX54 (NESG ID: PgR37) contains disordered regions. Cloning and expression analysis of the full-length protein in our bacterial expression system results in no detectable expression, supporting the disorder prediction based on the fact that proteins with significant disorder are often degraded in the *E. coli* cell. The construct optimization software generated alternative constructs based on the previously outlined criteria and the database of Protein Families that includes their annotations and multiple sequence alignments (DUF477) domain boundary (residues 59-182). These are depicted in Figure 1B. The two expression constructs comprised of residues 54 - 187 and 59 - 182 also did not express at detectable levels. However a slightly longer construct (residues 35-182) was highly expressed and soluble and ultimately allowed the structure of this protein to be solved by NMR (Protein Data Bank ID:2KW7). Interestingly the three-dimensional structure reveals the presence of two short helical regions between residues 38-48 and a β -strand for residues 54-56. These helices and the β -strand are tightly packed against each other and to other regions of the

protein. The loss of these interactions in the shorter constructs likely destabilizes the protein, leading to degradation in the expression host.

Alternative constructs generated by the Construct Optimization Software are reviewed by a bioinformatics expert and, if approved, are entered into the NESG Protein Laboratory Information Management System (PLIMS). This JAVA-based Oracle database provides a detailed protein production data model following lab activities on a step-by-step basis. PLIMS is a web-based application consisting of four main modules: (i) Target Registration & Management, (ii) Molecular Biology & Analytical-Scale Protein Expression, (iii) Large-scale Fermentation, and (iv) Protein Purification. PLIMS captures all pertinent information during the protein sample production process, interfacing where possible with robotics. It utilizes bar codes, personal digital assistants, and wireless technology. Key data from PLIMS is then uploaded to the internet-accessible NESG SPINE Structure Production Database (Bertone et al., 2001; Goh et al., 2003) to be shared across the consortium and with public databases, including the Protein Structure Initiative TargetDB (<http://www.sbkdb.org/>) (Chen et al., 2004).

The final step before wet laboratory work can commence is the design of polymerase chain reaction (PCR) primers. The DNA sequences for the proposed alternative constructs are generated by the PLIMS database, usually in 96-well format. These sequences are entered into the freely available web-based software Primer Prim'er software (http://www.nesg.org/primer_primer) for automated primer design (Everett et al., 2004). Vector specific PCR primer sets are designed to amplify and insert targeted regions into a vector of choice. The NESG has designed the “NESG Multiplex Vector Kit”, a series of vectors with a common multiple cloning site designed to minimize the number of non native residues while adding a short 6X-His tag (Acton et al., 2005). These expression vectors are available from the Protein Structure Initiative Materials Repository (<http://psimr.asu.edu/>). Although affinity tags are generally required for high-throughput purification protocols (Crowe et al., 1994; Sheibani, 1999) many commercial vector systems contain numerous non native residues (which are likely disordered) in their extended tags. Such large, disordered purification tags can produce large sharp peaks in the NMR spectrum that interfere with NMR studies. Primer Prim'er supports classical restriction endonuclease cloning, viral recombination cloning strategies such as Gateway Cloning (Invitrogen) (Hartley et al., 2000), as well as the InFusion (Clontech) based cloning most commonly used in the NESG sample production pipeline. It can also be easily modified to support other cloning strategies. In this task, Primer Prim'er generates ORF-specific primers with regions of vector overlap for use with the ligation independent cloning strategy as discussed in detail below. The software then organizes the primers into 96-well format, separating forward and reverse primers onto two different plates but in corresponding wells. Additionally, the user can array the primers in a variety of fashions, grouping primer sets on expected size of amplification (increasing, decreasing, staggered), template source (e.g., all *E. coli* targets grouped together), etc. (Everett et al., 2004). The arrayed primer sets are entered into PLIMS, which generates order forms for an oligonucleotide vendor.

2. Ligation Independent High-Throughput Cloning and Analytical Scale-Expression Screening

2.1 Ligation-independent cloning and automated vector construction

The first step in the high-throughput cloning is the procurement of template DNA for the PCR amplification of the coding sequence regions selected by the construct design process. Unlike oligonucleotide primers that are chemically synthesized and easily available from a variety of vendors at inexpensive rates, obtaining high quality PCR template has a number of issues. For prokaryotic targets, the NESG has organized a collection of genomic template for over 150 Reagent Genomes. In many cases, Whole Genome Amplification by Multiple Displacement Amplification has been used to generate template genomic DNA for our prokaryotic Reagent Genomes (Acton et al., 2005; Xiao et al., 2010). However, in recent years, we have had growing emphasis on production of human and other eukaryotic proteins. As *E. coli* does not have the splicing machinery to allow the use of eukaryotic genomic DNA as PCR template for the expression of genes containing introns, cDNA must be used.

Whole genome amplification by multiple displacement amplification—Although the complete genomic sequence of over 1200 prokaryotic organism has been elucidated, with even more in progress, genomic DNA preparations are commercially available for only a small fraction (~10%). Recent methods that predict success in expression, solubility, crystallization, and NMR spectral quality, based on primary sequence (Slabinski et al., 2007; Price et al., 2009) often identify a prokaryotic member of a given protein domain family as amenable to structural determination. However, genomic DNA preparations from these organisms are often not available, and extracting genomic DNA from these organisms can be problematic. To circumvent these issues and increase project efficiency by expanding our reagent genome list, we have implemented Whole Genome Amplification by Multiple Displacement Amplification utilizing phi29 DNA polymerase (Kvist et al., 2007; Lasken, 2007; Lasken, 2009), to produce microgram quantities of genomic DNA suitable for use as cloning template (Dean et al., 2002). Specifically, a small aliquot of freeze-dried cells from the ATCC (American Type Culture Collection), most sequenced prokaryotic strains are available from this resource, are used as template for Multiple Displacement Amplification. We have generated high molecular weight genomic DNA preparations for over 30 new prokaryotic genomes from organisms ranging from gut metagenomic bacteria (Gill et al., 2006; Qin et al., 2010) to extremophiles using this high-fidelity technique, greatly expanding the range of proteomes that we can target (Acton et al., 2010).

RT-PCR—Numerous cDNA sources are commercially available ranging from cDNA clone pools, full-length verified clones (length verified by sequencing), and full-length, fully-sequenced cDNA clones, such as the ORFeome collaboration clones (Open Biosystems) (Rual et al., 2004; Lamesch et al., 2007). Only the latter type of cDNA clone is flawless as a template source with respect to sequence, while the others may contain PCR based errors and polymorphisms. All individual clone libraries have logistical issues, they must be purchased (at considerable cost), archived, and rearranged, before use as PCR template. To circumvent these problems, we have developed protocols for reverse transcriptase mediated

generation of cDNA from PolyA RNA extracted from various eukaryotic organisms (RT-PCR). cDNA pools are generated from commercially available polyadenylated mRNA preparations from various tissues, cell types, and developmental stages (PolyA⁺ RNA, Clontech), including a considerable number of tumor cells and human cell lines. The PolyA⁺ RNA is incubated with oligo-dT or random primers and MMLV reverse transcriptase to generate cDNA pools, which are then combined and used as a common template that is added to each PCR reaction with target specific primers much like using bacterial genomic DNA. Although this approach may generate clones with polymorphisms (a problem also when using cDNA clone collections that are not full-length and fully-sequenced) we find it effective in terms of cost, amenability to high-throughput manipulations and PCR efficiency. Indeed, our PLIMS database indicates that 88% of GC rich (>59% GC content) and 96% of lower GC content RT-PCR amplification products are of the correct size. This strategy has proven successful for cloning genes suitable for protein sample production from *Homo sapiens*, *Bos taurus*, *Mus musculus*, *Rattus norvegicus* and *Arabidopsis thaliana* and from other eukaryotic organisms.

Vector construction by ligation-independent cloning—The NESG originally developed a restriction endonuclease/T4 DNA ligase-based 96 well cloning strategy, using our Multiplex Cloning Vector Set and a Qiagen BioRobot 8000 automated liquid handling device (Acton et al., 2005). The vector system minimizes the number of non native residues in the open reading frame while adding a 6X-His tag and common polylinker sequences for purification and cross compatibility, respectively. Although the vector set is comprised of 9 different vectors in three different reading frames, the majority of NESG cloned targets reside in one of three NESG-modified T7 expression vector derivatives: pET15_NESG, pET21_NESG or pET15TEV_NESG, coding for proteins with N- (MGHHHHHHSH-), C- (-LEHHHHHH), or N- (MGHHHHHHENLYFQSH-) 6X-His affinity purification tags respectively. The latter expression vector contains a TEV protease cleavage site for the removal of the 6X-His tag (Nallamsetty et al., 2004).

During Protein Structure Initiative-2, InFusion-based ligation-independent cloning was introduced to increase cloning efficiency and throughput. Like other ligation-independent cloning systems, InFusion is more efficient, less time consuming and requires less technical skill than ligase-dependent cloning (Aslanidis and de Jong, 1990; Haun and Moss, 1992). No vector modification is necessary allowing complete compatibility with the NESG Multiplex Expression Vectors, keeping the number of non native residues in the expressed construct to a minimum. The InFusion enzyme only requires the addition of a 15 base pair tail to each of the gene specific PCR primers for a given target ORF. As depicted in the lower center portion of Figure 2, these base pairs are identical to the 5' and 3' regions of the vector multicloning site respectively (Zhu et al., 2007); the 15 base pair region of overlap (identical double-stranded DNA sequence) at the 5' end is shown in gray, and the 3' end is shown in black, on both the vector and PCR amplification product (insert). The vector is cleaved at the region of overlap to produce linear vector DNA. The InFusion enzyme recognizes the identical ends of the two linear DNA strands (vector and insert) and through its strand displacement and exonuclease activity, promotes the pairing and resecting of the

DNA (Marsischky and LaBaer, 2004). Transformation of the resulting DNA complex into *E. coli* results in ligated circular DNA clones.

Vector preparation for InFusion cloning is nearly identical to restriction endonuclease cloning (Acton et al., 2005). To minimize non native residues the vector is cleaved at the outermost restriction sites in the multicloning site, NdeI, followed by XhoI restriction endonuclease digestion. The linearized vector DNA is purified by agarose gel electrophoresis, followed by gel extraction and normalized to 8 ng/μL.

High-throughput construction of expression plasmids—Having addressed important cloning related issues, alternative PCR template strategies, primer design, and the InFusion cloning approach provides the framework for a detailed description of the high-throughput cloning procedure. Figure 2 outlines each step of the high-throughput vector construction pipeline and illustrates the contributions of the PLIMS system and the BioRobot 8000 automation in our 96-well high-throughput cloning. A detailed protocol of the entire process can be downloaded (<http://www-nmr.cabm.rutgers.edu/labdocuments/proteinprod/index.htm>), including a description of custom Qiasoft 4.1 programs developed in house to perform the automated steps. Briefly, PLIMS-generated 96-well primer plates are provided by the vendor (Eurofin MWG Operon) at 50 μM concentration. Forward and reverse primers for each specific ORF (identical wells on two separate 96-well blocks) are placed on the BioRobot. An ABI thermocycler compatible 96-well PCR plate is chilled (4°C) on a temperature controlled BioRobot 8000 slot, and the eight-channel pipette head transfers 46 μL of an appropriate PCR reaction mix (dNTPs, high-fidelity thermostable DNA polymerase and buffer, template DNA, and nuclease free H₂O) to each well in a 96-well PCR plate. Multiple PCR reaction mixes are prepared for each source of template DNA (cDNA, prokaryotic genomic DNA, etc.) The BioRobot then transfers 100 pmol of the appropriate forward and reverse primers from the primer blocks into the corresponding well for each target in the PCR plate. The PCR plate is covered with a MicroAmp™ 96-Well Full Plate Cover (Applied Biosystems) and transferred to a Veriti® (Applied Biosystems) PCR thermocycler for amplification. Primer Primer generates primers with similar melting temperatures allowing thermocycling parameters to be standardized across the plate. Each cycle contains a 10 s, 94 °C melting step, a 30 seconds annealing step (50–55 °C), and a 3 minutes 68 °C elongation step (~1 minute per kilobase pair). An annealing temperature step increase after 10 rounds of amplification (57–60 °C) is included for the final 25 cycles to adjust for the increased stability derived from the added recombination sites base pairs (Xiao et al., 2010). Advantage® HF2 Polymerase (Clontech) is normally used for the initial PCR of most targets. This enzyme preparation has proven extremely robust, with very high fidelity. However, human target genes, as well as some prokaryotic strains added to our Reagent Genome List during Protein Structure Initiative-2, may have high GC-content. As the GC content increases, PCR amplification becomes increasingly more challenging. Alternative thermostable DNA polymerases (Advantage® GC 2 Polymerase; Clontech) have proven to be very useful for such GC-rich PCR amplifications. Various concentrations of DMSO are also evaluated to produce PCR products of the correct size and of sufficient quantity for high-throughput cloning. Although higher error rates will occur with these GC-rich PCR conditions, efforts must be made (often opposing) to adjust buffer and annealing

temperature conditions to maximize fidelity while increasing the likelihood of obtaining amplification product.

Following thermocycling, 10X DNA Loading Dye is added to each well of the PCR plate using the BioRobot 8000. A Matrix EXP 8-Channel Pipette (Thermo Scientific) with adjustable tip spacing is used to load the PCR products on a 2% agarose gel for separation and visualization. The results are documented by Alpha Imager (Cell Biosciences) documentation, entered into the PLIMS database, and successful PCR amplifications are identified (correct size and acceptable yield). DNA fragments for the successful PCR reactions are excised from the gel with a SafeXtractor (5 Prime) and transferred to the appropriate well of a 96-well S-Block (Qiagen). The BioRobot8000 performs an automated 96-well gel extraction with a protocol developed using Buffer QG from the Qiagen Gel Extraction Kit and a QIAquick 96-well column PCR Cleanup plate. An aliquot (1 μ L) of the resulting purified PCR products (~40 μ L/well) is transferred to the appropriate well of a fresh PCR plate. A reaction mix for 96 well ligation-independent cloning is then prepared by adding 96 μ L of the above treated vector (8 ng/ μ L) with 288 μ L of sterile deionized water. The reaction mix is used to rehydrate 24 wells from the Infusion Dry Down PCR Cloning Kit (96-well). A considerable cost savings is achieved by diluting the InFusion reaction 4-fold over vendor suggestion (4 reactions per well or 384 reactions per plate). The InFusion reaction mix is aliquoted to each well (4 μ L) and the plate covered and incubated for 15 minutes at 37 °C followed by 15 minutes at 50 °C, the reaction mix is then flash frozen (−80 °C). The plate is thawed on ice and the 96 InFusion reactions (resected and paired vector and insert) are transformed into *E. coli* cells, using a 24-well format robotic transformation procedure. A single microliter of the ligation-independent cloning product is robotically transferred to the corresponding well of a fresh 96-well PCR plate containing 10 μ L of XL-10 Gold® Ultracompetent Cells (Agilent) chilled at 0 °C on the robot deck. Following incubation for 30 minutes at 0 °C, a manual heat shock step (one minute at 42 °C) is performed, SOC (100 μ L) is added to each well, and the plate is incubated at 37 °C for 1 h. The robotic 8-channel pipette head transfers the entire content of each well to a corresponding well in one of four 24-well blocks. The platform shaker distributes the transformation reactions via 5–10 (3-mm-diameter) glass beads over 2 mL of Luria Broth (LB) medium/Agar with ampicillin (100 μ g/mL) contained in each well. Following overnight incubation at 37 °C, two colonies per ORF are harvested for colony PCR, using primers flanking the multiple cloning site. Colonies arising from empty vector are rare since the InFusion enzyme does not have ligase activity, and self-ligation by host enzymes appears inefficient with the minute overhangs produced by restriction digest. The colony PCR reactions are loaded on a 2% agarose gel and the results are documented in the PLIMS database. Correct transformants are transferred to a PLIMS directed well of a 48 well block (Qiagen) containing 2 mL of LB/ampicillin and grown overnight on a platform shaker (37 °C, 210 rpm). Cells are harvested by centrifugation (3000g-force for 10 min) and the media is discarded. Plasmid DNA is isolated using a completely automated QIAprep 96 Turbo Miniprep Kit and BioRobot 8000 procedure. Both the overnight culture and miniprep DNA are archived in an NESG Reagent Repository. Using the InFusion ligation-independent cloning method we have cloned over 20,000 constructs of some 9,000 unique protein targets (multiple alternative constructs per target) into pET expression vectors.

2.2 Analytical scale expression screening

We have developed a microtiter plate based high-throughput analytical scale protein expression system to screen the numerous expression constructs produced by high-throughput cloning. The overall goal of this assay is to identify protein constructs or homologues that will produce highly-expressed, soluble samples for structural analysis when produced by preparative fermentation (1-3 liters) in minimal media suitable for isotope enrichment or selenomethionine labeling. Although nearly one hundred new constructs are produced in each cloning set, only 30-50% of the clones (depending on the source of the protein) will express soluble protein at the levels needed for NMR studies. Such high attrition rates exclude performing large-scale fermentation on every construct due to high costs and inefficiency. We have therefore developed the plate-based strategy to evaluate expression (*E*) and solubility (*S*) in a high-throughput fashion, while maintaining the highly-aerated growth conditions found in later fermentation efforts.

An outline of this procedure is presented in Figure 3. It starts with transformation into a codon-enhanced *E. coli* expression strain. For each construct at least two isolates are pursued. Often this identifies a clone with more favorable characteristics. Typically, codon-enhanced strains BL21(DE3)-Gold (Agilent), harboring the pMgK plasmid that encodes for rare tRNA codons of Arg (AGA, AGG) and Isoleucine (ATA), or BL21-CodonPlus(DE3)-RIPL (Agilent) that additionally supplies rare tRNA codons for Pro (CCC) and Leu (CTA), are used. Briefly, a robotic transformation protocol is performed, 1 μ L of each expression plasmid is transferred to the corresponding well of a PCR plate (on ice) containing 10 μ L of competent cells. Following a 30 minute incubation period the plate is transferred to a PCR thermocycler for a 1 minute heat shock at 42 °C. The cells are transferred to the corresponding well of one of four Falcon Multiwell 24 well tissue culture plates (Becton-Dickinson) containing 2 mL of LB agar (with appropriate antibiotics and 0.5% glucose) and 5-10 (3 mm) borosilicate glass beads per well. The latter is used to distribute the cells much like the transformation step during ligation-independent cloning. Following overnight growth, individual colonies are inoculated into the corresponding well of a 96-well S-block (Qiagen) containing 0.5 mL of selective LB medium per well. The plate is covered with AirPore Tape (Qiagen) and incubated for 6 h at 37 °C on an Innova platform shaker (New Brunswick Scientific) rotating at roughly 210 rpm. A simple microplate freezer rack can be modified and attached to the shaker platform to provide a low cost method to simultaneously secure up to 24 deep well microplate blocks for shaking incubation (see Figure 3). Following the allotted time, the BioRobot transfers 10 μ L of each well into the corresponding well of a fresh 96-well S-block containing 0.5 mL (per well) of MJ9 minimal media (Jansson *et al.*, 1996). The block is covered with AirPore Tape and incubated overnight (37 °C at 210 rpm). We have found that growth under minimal media conditions differs significantly from growth in rich media (e.g., LB), often affecting expression and solubility behavior. Therefore, analytical scale expression utilizes the same minimal media as preparative-scale fermentation in an effort to maximize reproducibility between the two expression scales. The MJ9 minimal media allows for either isotope enrichment or selenomethionine labeling in preparative-scale fermentation.

Following overnight incubation, the BioRobot performs a 1:20 dilution of the saturated growth into one of four 24-square-well blocks (10 mL maximum volume/well) containing 2 mL of MJ9 media, preserving well assignment. Each block is covered with Airpore Tape and incubated at 37 °C (210 rpm) until mid-log phase (2–3 h growth, 0.5–1.0 OD₆₀₀ units) is reached. The 24-well blocks are cooled and isopropyl thiogalactopyranoside (IPTG, 1 mM final concentration) is added to each well to induce expression. The block is resealed with fresh AirPore Tape, and incubated overnight at 17°C with shaking at 210 rpm. The low temperature incubation often aids in producing soluble proteins (Shirano and Shibata, 1990), while the vigorous shaking with gas permeable tape allows for greater aeration rates like those that we obtain in our Midi-scale fermentor (described below) or large-scale fermentation in baffled flasks. Following overnight induction, cells are harvested by centrifugation; the pellets are resuspended in Lysis Buffer (50 mM NaH₂PO₄, 300 mM NaCl, 10 mM 2-mercaptoethanol) and robotically transferred to a 96-well round bottom plate (Greiner). A chilled 96-probe sonicator (Qsonica, LLC.) is used for cell disruption, with a 12 minute cycle time (30 sec bursts at 18 watts followed by 30 sec cooling periods). Total and soluble portions of the cell lysate are then analyzed by SDS-PAGE (see Figure 3 for details). Expression (*E*) and solubility (*S*) are scored, each on a scale of 0 (none) to 5 (max); i.e. the $E \times S$ (or *ES* value) ranges from 0 to 25. All data is documented in the PLIMS system, and 96-well glycerol stocks of the constructs in the expression strain are archived in the NESG Reagent Repository.

Based on the expression and solubility levels obtained from subsequent large-scale (1 – 3 L) fermentations we have defined $ES > 11$ as a “usability score”. Expression constructs meeting this criterion generally provide enough material (5 – 50 mg / L) from preparative scale fermentation and protein purification to allow high-throughput structural determination techniques. In spite of the challenges in producing similar growth conditions in plate format and 1 – 2 mL volume versus 1000 mL volumes, after considerable efforts, we have achieved 90% agreement between the different expression scales. This allows accurate prediction of large-scale expression yields.

3 Midi-Scale Protein Expression and Purification

A retrospective analysis of the NESG expression constructs produced during the second phase of the Protein Structure Initiative indicates that roughly one-third of all expression constructs possess expression and solubility levels (i.e., $ES > 11$) consistent with the requirements for large-scale fermentation. However, a large fraction of these samples prove intractable at later steps in the protein production pipeline. For example, roughly one-half of the proteins produced in the last two years are not monodisperse in solution and contain some level of aggregation. Analysis of NESG data indicates that crystallization success rates are dramatically increased more than 10-fold for monodisperse protein samples in comparison with those polydisperse or aggregated (Price et al., 2009). Protein aggregation is also deleterious for NMR studies, often resulting in protein precipitation or negatively affecting NMR spectra. Approximately one-half of protein samples prepared for HSQC screening provide “Good” or “Promising” HSQC spectra and are amenable to structural determination by NMR.

Since about one-half of purified proteins have been intractable for structural studies due to their biophysical properties, we have developed a high-throughput Midi-scale Protein Production Pipeline which provides a biophysical characterization of small quantities of protein constructs before investing in more labor-intensive large-scale expression and purification. This system, outlined in Figure 4, utilizes (i) a GNF Airlift Fermentation System (GNF Systems) with O₂ aeration allowing for 96 simultaneous high cell density fermentations; (ii) a His MultiTrap HP plate (GE Healthcare) for 96 parallel immobilized metal affinity chromatography (IMAC) purifications; and (iii) Zeba™ 96-well desalting spin plate (Thermo Scientific) for buffer exchange. Typical yields of 0.2 - 1.0 mg of protein per 60 mL fermentation are achieved, which is sufficient for a series of analytical protein chemistry steps including: (i) aggregation screening by analytical gel filtration with static light scattering, (ii) homogeneity analysis by SDS-PAGE analysis, (iii) target validation by MALDI-TOF mass spectrometry, (iv) concentration determination by a NanoDrop ND-8000 spectrophotometer, and (v) 1D ¹H-NMR screening using a 1.7-mm micro cryo NMR probe (35 uL sample volume) (Rossi et al., 2010). Identification of aggregated/polydisperse proteins and proteins exhibiting poor quality 1D ¹H-NMR spectra avoids scale-up of intractable protein targets, and allows us to concentrate finite resources for isotope labeling and protein purification on targets amenable to NMR studies or crystallization screening, greatly increasing project efficiency

3.1. Midi-scale fermentation with the GNF airlift fermentation system

To produce sufficient quantities of protein for biophysical characterization in a high-throughput fashion we have adapted an Airlift Fermentation System, developed by the Genomics Institute of the Novartis Research Foundation (GNF), to our Midi-scale pipeline. This system uses 96 (100 ml) test tubes, each with a fermentation capacity of ~60 mL of media. A common manifold with 96 canulae is used to deliver oxygen to each tube for both cell growth, and to provide agitation for cell suspension and nutrient mixing. Temperature regulation is controlled using a water bath with a refrigerated/heated water circulator (VWR Scientific). Using rich TB media (Peti et al., 2005), this system routinely reaches cell densities in the range of 15-20 OD₆₀₀ units, corresponding to a quarter of the final cell mass obtained from 1 L of our large-scale protein expression in minimal media (3-5 OD₆₀₀ units). Briefly this procedure starts with PLIMS generating a target pool of constructs with high *ES* values (> 11). After selecting 96 targets, robotic transfer from the appropriate glycerol stock to a PLIMS-directed well in a 96-well S-block is performed. Each well contains 500 µL of TB media with ampicillin and kanamycin and is then covered with AirPore tape. Following incubation (37 °C, 210 rpm) for six hours, 100 µL from each well is transferred to a 100 mL test tube in the corresponding position of the GNF fermentor. Each tube contains 3 mL of TB, Antifoam Y-30 Emulsion (Sigma) and appropriate antibiotics. The rack is covered with aluminum foil and placed on an Innova shaking platform. Following overnight growth at 37 °C, 57 mL of fresh TB/antifoam/antibiotics are added to each of the 96 tubes. The air intake manifold is inserted into the tube rack and the entire system is relocated into a water bath preheated to 37 °C. Using the manifold and its canulae, 100% oxygen is distributed to each well at a flow rate of ~3.5 cfm. We have found that the dual functioning canulae, providing both oxygenation and lift (large bubble size), necessitate a high percentage of oxygen addition for the greatest yield. This is potentially dangerous, and strong system ventilation

must be used for safety. When OD₆₀₀ reaches 5-6 units, each tube receives IPTG (1 mM final concentration) and Antifoam Y-30 Emulsion through ports in the Airlift Fermentation System manifold. Concurrently, the water bath temperature is decreased to 17 °C using the refrigerated water circulator. Following 16 hours of incubation at this temperature and aeration with 100% oxygen, an aliquot is taken from each well to assay final cell density and for SDS-PAGE analysis of expression and solubility levels. The resulting data is documented and processed in the PLIMS database. The remaining contents of each tube are transferred to an appropriately labeled (PLIMS generated well position and bar-code) 50 mL conical tube, and centrifuged for 20 minutes at 6000 g-force. The media is poured off and the pellets are flash frozen and stored at -80 °C.

3.2 Ni²⁺-affinity protein purification using 96-well IMAC plates

The Midi-scale fermentation system produces sufficient protein levels for carrying out numerous biophysical characterization techniques. To purify these proteins to levels suitable for these techniques we have also developed a Ni²⁺-based IMAC step in plate format. This takes advantage of the 6X-His tag incorporated into each construct. Briefly, the cell pellets are thawed and resuspended in Lysis Buffer containing 1 × Cell Lytic B (Sigma), 500 µg/ml lysozyme (Sigma), 100 units/ml RNase (Sigma), 100 units/ml DNase (Sigma), 40 mM imidazole and Complete Protease Inhibitor Cocktail (Roche). The resuspension mix is incubated for 30 minutes at 37 °C. Each 50 mL tube is then centrifuged at 3,000 rpm (1600 g-force) for 30 minutes to clear the cell debris. 2 mL of each resulting supernatant is transferred to an empty 2.2-ml deep-well plate (Qiagen S-block). A Liquidator96 (Rainin) is used to transfer 400 µL from each well to the corresponding well of a His MultiTrap™ HP 96-well plate (GE Healthcare) equilibrated with Lysis Buffer. The plate is centrifuged for 4 minutes at 100 g-force and the flow through is discarded. This process is repeated, loading the entire 2 mL of cell lysate into each respective well. The IMAC plate is subjected to three wash steps consisting of 500 µL of Lysis Buffer (per well) containing 40 mM imidazole (pH 7.5) followed by centrifugation at 500 g-force for 2 min. Proteins are eluted from the Nickel Sepharose with Lysis Buffer containing 300 mM imidazole (pH 7.5) per well. 75 µL of this solution is added to each well and followed by incubation at room temperature for 10 min. The step is completed by centrifugation at 100 g-force for 4 min. The entire volume of Ni²⁺-affinity purified proteins are then immediately transferred to an equilibrated Zeba™ 96-well desalting spin plate (Thermo Scientific). Following centrifugation for 2 minutes at 1000 g-force, the proteins are ready for biophysical characterization.

3.3 Biophysical characterization of the midi-scale generated proteins

(i) Target validation by MALDI-TOF mass spectrometry—At this point in the NESG Protein Production Pipeline, DNA sequence verification of expression constructs has generally not been performed. Therefore, we have implemented a quality control step using MALDI-TOF mass spectrometry (MS) to identify the protein molecular weight of the protein produced by the expression vector in each well. Samples are prepared by mixing 1 µL of the protein sample from each well with 10 µL of sinapinic acid matrix solution (10 mg/ml sinapinic acid in 50% acetonitrile/50% 0.1% TFA) in the corresponding well of a PCR plate. 1 µL of this solution is transferred to the appropriate well position on an Opti-TOF Sample Plate (AB SCIEX) and spectra are collected for each protein spot

(corresponding to a well position) on a MALDI-TOF/TOF (4800 Plus MALDITOF/TOF™ Analyzer, AB SCIEX) in single TOF mode. The spectrum of each well is compared to the expected size of the purified protein; species differing from their expected mass by greater than 500 Daltons likely represent invalid, processed, or proteolyzed targets. Biophysical analysis continues on those samples with aberrant molecular weight, however, these expression construct and purified protein are subjected to DNA sequence and liquid chromatography mass spectrometry analysis for verification of the cloned DNA sequence and protein product, respectively. These mass spectrometry data are all archived in the SPINE database (Goh et al., 2003).

(ii) NanoDrop and SDS-PAGE for assaying protein concentration and homogeneity—To measure protein concentration in standard non-high-throughput applications, protein samples are transferred to cuvettes or capillary tubes for spectrophotometric measurement of absorbance at 280 nm. In many high-throughput projects, absorbance measurements are performed on solutions in microtiter plates. However, we have found that neither of these methods is well suited for the Midi-scale pipeline since the samples are of limited volume and large in number. To circumvent these issues, we have incorporated a NanoDrop™ 8000 spectrophotometer into the NESG Protein Production Pipeline. This instrument accurately measures absorbance values of sample volumes < 2 µL without dilution. Briefly, a Multichannel Pipette (Rainin) is used to simultaneously transfer single columns (8 samples, e.g., A1-H1) of the purified proteins from the PCR plate to a linear array of 8 pedestals. The pedestals are arrayed with standard 96-well plate spacing, each with optics for absorbance measurement. Using this system a full plate of 96 samples can be measured in less than six minutes. The protein concentration in each well is calculated automatically using its respective extinction coefficient (calculated by the PLIMS system). The concentration of each protein is recorded in the SPINE database and used in the Aggregation Screening analysis (described below), in the preparation of NMR samples, and to determine process yield.

An important quality control step in protein purification is to assess the homogeneity of the preparation. For the larger proteins in the NESG X-ray crystallography protein production pipeline we utilize a LabChip® 90 system (Caliper) for this task. Although this system is ideal for high-throughput procedures, it is technologically limited to proteins greater than 12 kDa and cannot be used for some of the smaller proteins produced in the NMR protein production pipeline. For these smaller proteins, homogeneity is accessed by SDS-PAGE analysis. NuPage® Novex® Bis-Tris Mini-Gels (Invitrogen) are utilized for ease of sample preparation, excellent resolution, long-shelf life, and high reproducibility. Samples are prepared by aliquoting 2 µL of the purified protein into a fresh PCR plate containing 5.5 µL of deionized water, 0.5 µL of DTT (1 mM) and 2.5 µL of NuPage LDS Sample Buffer per well. The samples are mixed and heated at 70 °C for 10 min. Gels are loaded using an eight-channel expandable pipette (Matrix EXP Pipette, Thermo Scientific) that reduces from normal 96 well spacing to that of the SDS-PAGE gel. Gels are fixed and stained using GelCode Blue (Thermo Scientific), documented (AlphaImager™, Cell Biosciences™) and archived in SPINE.

(iii) Aggregation screening to detect protein mass distribution—Numerous studies have shown that proteins that are monodisperse in solution are more likely to produce diffraction quality crystals than polydisperse or aggregated samples (Ferre-D'Amare and Burley, 1994; Ferre-D'Amare and Burley, 1997; Klock et al., 2008; Price et al., 2009). Roughly one-half of NESG structures are determined by X-ray crystallography. Therefore, to increase project efficiency an Aggregation Screening step was implemented to characterize purified proteins by analytical gel filtration followed by multi-angle static light scattering. This allows for accurate measurement of the distribution of oligomers and/or aggregates in a protein sample. Although this system was originally implemented for X-ray crystallography samples, probing the oligomerization state of NMR samples is also valuable for assessing the methods to be used in NMR studies; e.g., monomer and homodimers require different data collection and analysis strategies, and higher order multimers may be beyond the size constraints for high-throughput NMR structural determination. These Aggregation Screening data also predict NMR sample stability (proteins with significant aggregation may precipitate before a full data set is collected). Aggregation Screening is carried out using an Agilent 1200 series HPLC system with an 8 mm × 300 mm Shodex KW-802.5 HPLC size-exclusion column (Showa Denko K.K.), connected in line with a miniDAWN TREOS detector (Wyatt technologies) and a Optilab rEX Refractometer (Wyatt Technology). Protein samples are loaded from the automated 96-well sample changer and injected into the HPLC system equilibrated in a candidate NMR buffer (see section 3.3.iv). Following separation by size exclusion, the light scattering properties of each successive protein species in solution is simultaneously measured at three different angles (45°, 90°, and 135°) and their refractive index is detected. The analysis of the light scattering and refractive index data provides the shape-independent weight-average molecular mass of each species and their relative distributions. As shown in Figure 4, the light scattering trace for the NESG human protein target HR3580C indicates peaks corresponding to monomer (M), dimer (D), higher oligomers (O), and aggregates (A) of the protein (panel marked Aggregation Screening). The bottom trace shows the refractive index indicating that the majority of mass is contained as a monomer. However, analysis of the data indicates that ~75% of the mass is monomeric (22.9 kDa) in nature. Although generally monomeric and therefore within the size-limits of high-throughput NMR assignment and structural analysis, the significant amount of aggregated protein observed suggests that further buffer optimization, construct optimization, or other “salvage” efforts are required before promotion to large-scale fermentation and purification.

(iv) High-throughput micro cryo probe screening by 1D ¹H NMR—The NESG Microscale Protein NMR Sample-Screening Pipeline has been described in detail (Rossi et al., 2010). It takes advantage of NMR microprobe technology and their inherent ability to function with relatively low quantities of protein. Typically, 10-200 micrograms of protein in a volume of 35 µL is sufficient for screening with a Bruker 600 MHz TXI 1.7-mm micro cryo probe. This probe is well suited for the yields generated by our Midi-scale Pipeline. Briefly, the 96 purified proteins are buffer exchanged into an appropriate NMR buffer using a Zeba™ 96-well desalting spin plate (Thermo Scientific) as described in Section 6.1. The initial NMR buffer is selected based on the isoelectric point (pI) of the protein; typically 20 mM MES, 100 mM NaCl, 5 mM CaCl₂, 10 mM DTT, 0.02% NaN₃ at pH 6.5, or

alternatively 20 mM ammonium acetate, 100 mM NaCl, 5 mM CaCl₂, 10 mM DTT, 0.02% NaN₃ at pH 4.5. Aliquots are then transferred to 1.7 mm SampleJet Tubes (Bruker) using a Gilson 215 Liquid Handler. Only 1D ¹H NMR spectra are collected, since the rich TB broth in the Midi-scale Pipeline does not allow for isotope enrichment. However, this screen can detect dispersion of amide protons and upfield-shifted methyl protons, indicative of aromatic and methyl stacking (folded protein core). Protein constructs exhibiting tractable traits from the Midi-scale biophysical screening (correct molecular weight, monodisperse monomers or low molecular weight dimers, disperse amide and methyl proton resonance frequencies) are more than likely amenable for structure determination by NMR. These are subsequently scaled up for fermentation and purification with isotope enrichment.

4. Preparative Scale Fermentation

The recent advent of NMR microprobe technology has allowed for structural determination using relatively minute amounts (< 100 µg) of protein sample (Aramini et al., 2007). However, microcryo NMR probes have lower sensitivity than 5 mm cyro probes, which utilize significantly greater amounts of protein (>5 mg). We have designed our process for preparative-scale or large-scale protein expression to produce these yields while optimizing cost and throughput, and incorporating the flexibility to use essentially the same pipeline to produce samples for both NMR studies and X-ray crystallography. This strategy is based on parallel fermentations in 2.5-liter baffled Ultra Yield™ Fernbach flasks (Thomson Instrument Company) with agitation on low cost platform shakers (Innova 2300, New Brunswick Scientific) housed in controlled temperature rooms (Acton et al., 2005). The Ultra Yield™ flasks, with advanced 6-baffle design, are capable of supporting high cell density (Brodsky and Cronin, 2006), while their compact design allows for up to 15 parallel fermentations per platform shaker. Together with buffer- and vitamin-enriched MJ9 minimal media (Jansson et al., 1996), this system of flasks and shakers achieves cell density and protein expression levels consistent with the needs of a high-throughput structural genomics pipeline. MJ9 media allows for enrichment with ¹⁵N, ¹³C, and/or ²H isotopes for NMR structural studies, or with selenomethionine for single and multiple anomalous diffraction X-ray crystallography (Hendrickson, 1991). For most NMR studies undertaken by the NESG, an NC5 (100% ¹⁵N, 5% ¹³C) sample is first produced for 2D ¹H-¹⁵N HSQC screening (Rossi et al., 2010). NC5 samples are fermented in MJ9 media supplemented with uniformly (U)-¹⁵NH₄ salts as the sole nitrogen source while the glucose (sole carbon source) is 95% natural abundance and 5% (U)-¹³C enriched. The NC5 samples are also used for stereo-specific assignment of isopropyl methyl groups (Neri et al., 1989). For targets providing good quality HSQC spectra, (U)-¹³C, U-¹⁵N-enriched proteins are then produced using MJ9 media supplemented with (U)-¹³C glucose and (U)-¹⁵NH₄ salts as the sole source of carbon and nitrogen, respectively.

The fermentation process (outlined in Figure 5) begins with the PLIMS database generating a pool of protein expression constructs that pass analytical scale (*ES*>11) and Midi-scale biophysical characterization, and meet variety of project-defined criteria (e.g., structural uniqueness). The appropriate glycerol stock plate and well position for each selected construct is reported (each plate has a unique bar code identification), and the plates are retrieved from -80 °C storage. An aliquot from each selected well is robotically transferred

to 500 μ L of LB media with ampicillin and kanamycin or chloramphenicol [BL21(DE3)-Gold pMgK or BL21-CodonPlus(DE3)-RIPL, respectively] in an S-block and incubated for six hours at 37 °C. An aliquot (50 μ L) of this rich media preculture is then used to inoculate 50 mL of MJ9 minimal media in a 250 mL flask and incubated overnight at 37 °C on an Innova shaker (250 rpm). The entire volume of overnight culture is then used to inoculate a 2.5-liter Ultra Yield™ flask containing 1.0 liter of MJ9 (preheated to 37 °C) supplemented with the appropriate isotopes and selective agents. The cultures are then incubated at 37 °C (shaking at 250 rpm) until the OD₆₀₀ reaches of 0.6–0.8 units. The 2.5-liter Ultra Yield™ flasks containing the cultures are then transferred to an Innova 2300 Shaker in a 17 °C constant temperature room. Following equilibration at this temperature (~10 min), protein expression is induced with IPTG (1 mM final concentration). Incubation with vigorous shaking (250 rpm) at this temperature continues overnight with a total induction time of 16 hours. Before harvest, 250 μ L of Antifoam Y-30 Emulsion is added to each flask to disperse any accumulated bubbles that may interfere with harvest. Before any further manipulations, a 500 μ L aliquot of cells from each culture is transferred to a microcentrifuge tube for (i) determining final cell density (OD₆₀₀), which typically range from 3 to 5 O.D. units, (ii) SDS-PAGE analysis of expression (E) and solubility (S), and (iii) quality control by DNA sequence analysis. The remaining culture is transferred into a HarvestLine System Liner (Beckman Coulter) contained in a specialized 1000 mL polycarbonate centrifuge bottle and cap (J-Lite PC-1000, Beckman Coulter). The cells are harvested by centrifugation at 9000 g-force for 30 minutes using an Avanti J-26 XP centrifuge (Beckman Coulter). Following centrifugation, the supernatant is discarded and the HarvestLine System Liner bags are removed from the centrifuge bottle and archived at –80 °C in a PLIMS directed location. SDS-PAGE analysis of expression and solubility is performed in a manner similar to Analytical Scale Expression (Figure 5). The fermentation data is uploaded in the PLIMS database for archival, including automated gel labeling. A subset of this information is transferred to the web-based SPINE database (Bertone et al., 2001; Goh et al., 2003) for web-based access across the NESG consortium and/or with the public databases

5. Preparative-Scale Purification

One of the greatest challenges in producing highly homogenous protein preparations (suitable for NMR structure determination and/or crystallization) in a high-throughput manner is the diverse biophysical characteristic of the targets themselves. The addition of a short 6X-His tag (Porath, 1992) utilized by the NESG “Multiplex Vector System” addresses this issue by allowing the same affinity purification to be utilized across our diverse target set. The 6X-His tag and the associated Ni²⁺-based IMAC along with automated four-module ÄKTAexpress systems (GE Healthcare) serve as the foundation of the NESG purification pipeline. Although the IMAC purification achieves a high degree of purity for most bacterially overexpressed proteins, a size exclusion chromatography step is performed to produce the purity levels necessary for crystallization and NMR studies. Each module in the ÄKTAexpress system is fitted with four separate HisTrap HP columns (GE Healthcare) and one size-exclusion column per module allowing four separate completely automated affinity/gel filtration two-step purifications, or 16 total purifications per each four-module system, in less than twelve hours.

Briefly, the SPINE database generates a pool of fermented targets available for purification and their corresponding -80°C storage. The database also reports associated protein target characteristics critical for the purification effort including protein molecular weight with tag for the exact isotope enrichment strategy, protein isoelectric point and extinction coefficient among others. The centrifuge bags of targets selected for purification are thawed on ice, 30 mL of Binding Buffer (50 mM Tris-HCl, 500 mM NaCl, 40 mM Imidazole, 1 mM TCEP and 0.02% NaN_3 , pH 7.5) containing a protease inhibitor cocktail (Complete Protease Inhibitor Cocktail, Roche) is pipetted into each bag and the pellet is resuspended by simple hand manipulation of the cell paste. The bag contents are then transferred to a pre-chilled stainless steel 125 mL beaker (Vollrath®) and sonicated in an ice water bath using a Dual Horn $\frac{3}{4}$ " probe (Qsonica, LLC.) for ten minutes with 30 second bursts followed by 30 second cooling periods (30 seconds on/30 seconds off). The cell debris is cleared by centrifugation at 27,000 g-force for 40 min, followed by filtration (0.2 μm). The supernatant is then loaded onto one of 16 positions on the ÄKTAexpress system and a two-step automated purification protocol is performed using the preinstalled default settings (AF-GF). Specifically, a HisTrap HP IMAC column (5 ml) and Superdex 75 26/60 gel filtration column are run in a linear series. The 6X His-tagged proteins are eluted from the HisTrap column using 5 column volumes of Elution Buffer (50 mM Tris-HCl, 500 mM NaCl, 500 mM imidazole, 0.02% NaN_3 , pH 7.5) at a flow rate of 4 mL per min. The ÄKTAexpress system monitors the elution profile (A_{280}) and collects major peaks into internal storage loops, which are then automatically injected onto the Superdex 75 gel filtration column equilibrated in one of two NMR buffers: (i) 20 mM MES, 100 mM NaCl, 5 mM CaCl_2 , 10 mM DTT, 0.02% NaN_3 at pH 6.5, or alternatively (ii) 20 mM ammonium acetate, 100 mM NaCl, 5 mM CaCl_2 , 10 mM DTT, 0.02% NaN_3 at pH 4.5, based on the protein's theoretical pI. The elution is monitored by A_{280} detection and major peaks emanating from the gel filtration column are collected in 96 well blocks. The resulting fractions are analyzed by SDS-PAGE and pooled followed by concentration using Amicon ultrafiltration concentrators (Millipore). Molecular weight validation by MALDI-TOF mass spectrometry, homogeneity analysis by SDS-PAGE, aggregation screening by analytical gel filtration with static light scattering, and concentration determination is performed on each sample. The NMR sample preparation is completed with the addition of Complete Protease Inhibitor Cocktail (Roche), 10% $^2\text{H}_2\text{O}$ and 50 μM DSS (4,4-dimethyl-4-silapentane-1-sulfonic acid) as an internal NMR reference (Markley et al., 1998).

For NMR microprobe data collection, aliquots (35 μL) are transferred to 1.7-mm SampleJet Tubes (Bruker), using a Gilson 215 Liquid Handler. For samples destined for data collection using 5-mm probes, 300 μL aliquots are transferred into 4- or 5-mm Shigemi tubes. The ÄKTAexpress purification trace, MALDI-TOF, Aggregation Screening, sample concentration, and other data are archived into the SPINE database. SPINE also serves to direct and track shipping of samples to NESG NMR spectroscopists around the country for data collection and structural determination. The database coordinates this effort with bar code based registration of shipment tubes and automatically tracks shipments through the FedEx database.

6. Salvage Strategies

Depending on the source of the protein target, up to 70% of expression constructs will fail to express soluble proteins at a level consistent with preparation of structurable samples. Many of these are high value targets. Therefore alternative strategies and expression hosts have been explored in an attempt to rescue these targets. In addition, many proteins proceed through the pipeline providing samples that are nearly structurable, but are not quite suitable for high-throughput structural determination techniques. The NESG Protein Production Pipeline has explored and developed “salvage” techniques to enhance the behavior of these proteins in attempts to provide samples more amenable to NMR studies. Some of the most successful salvage strategies are described below.

6.1 NMR buffer optimization

During Protein Structure Initiative-2, roughly one half of all proteins purified for NMR screening produce 2D ^1H - ^{15}N HSQC spectra that were scored as “Good” or “Promising” (Rossi et al., 2010; Snyder et al., 2005). The former are sufficient for high-throughput NMR structural determination. However the “Promising” spectra are marginal in quality and often cannot be used for resonance assignment or structural analysis. In the case of proteins with “Good” spectra, a significant proportion of these proteins will suffer sample stability issues (e.g. slow precipitation) in the time frame of complete NMR data acquisition (4-10 days). Although targets are screened in one of two different buffers, based on avoiding close proximity to the estimated protein pI, neither of these buffer conditions may be optimal. We have found that alternative buffer conditions can often improve NMR spectral quality, reduce aggregation and sample precipitation, and improve sample stability (Rossi et al., 2010). To screen for conditions that promote stability or better quality spectra we have implemented a high-throughput Buffer Optimization procedure. Purified proteins with good spectra and problematic precipitation or “Promising” spectra are thawed on ice. For each row of a ZebaTM 96-well desalting spin plate (Thermo Scientific), the corresponding microcolumns are first washed with one of 12 alternative buffers (e.g. listed in Table 1), by centrifuging the plates for 2 minutes at 1000 g-force to remove the storage buffer. Then, each microcolumn is equilibrated with the corresponding buffer by eluting with 250 μL of buffer transferred from a 12 partition Microplate Reservoir (Seahorse Bioscience) using a 12-channel pipette, followed by centrifugation for 2 minutes at 1000 g-force. This step is repeated three times until each well has received 1000 μL of the appropriate buffer. A candidate-purified protein is then transferred (~ 70 μL) to each well to sample all twelve conditions. Buffer exchanged samples are collected following centrifugation for 2 minutes at 1000 g-force. The eluate containing the buffer-exchanged protein is then loaded into 1.7-mm NMR microprobe tubes using a Gilson 215 Liquid Handler. Each tube is stored at room temperature and scored by visual inspection for precipitation after 10 days.

1D ^1H NMR spectra with solvent presaturation are acquired for each of the buffer conditions, including the slowly precipitating samples. 2D ^1H - ^{15}N -HSQC spectra are acquired only on samples that qualify as “well folded” based on the dispersed amide protons and the upfield-shifted methyl protons in the 1D spectrum. The 2D ^1H - ^{15}N -HSQC spectra under different buffer conditions are then overlaid and assessed. The scoring is now based

on the solubility as well as “foldedness”. It has been observed that certain buffer conditions that provide excellent solubility provide spectra characteristic of partially or totally unfolded proteins. It is important to note that the presence or absence of some precipitation is not correlated with spectral quality, as some precipitation may occur without significant loss of NMR signal. The buffer conditions providing the best quality 2D HSQC NMR spectra and possessing sufficient sample stability are identified, and future protein samples are prepared in this buffer. A more detailed description of sample preparation for buffer optimization has been previously described (Rossi et al., 2010). This same technique is also used for screening for ligands that bind to the protein, and to screen for ligands that can improve the quality of NMR spectra.

6.2 Construct optimization using amide hydrogen deuterium exchange with mass spectrometry detection (HDX-MS)

One of the keys to the NESG Construct Optimization Software (and the associated increase in project efficiency) lies in the identification of predicted disordered regions by the DisMeta server, and elimination of these residues from the protein construct. In spite of these efforts, expert analysis of the 2D ^1H - ^{15}N HSQC spectra generated by the NESG high-throughput NMR Microprobe Screening Pipeline can often identify regions of disorder in some proteins with “Good” or “Promising” spectra. Although it is possible in some cases to identify these regions following resonance assignment, this approach is time consuming. For this reason, we have implemented HDX-MS to experimentally identify these disordered regions in a high-throughput manner (Englander, 2006; Sharma et al., 2009; Woods and Hamuro, 2001). HDX-MS studies, outlined in Figure 6A, are based on the concept that backbone amide protons in disordered regions are solvent accessible and therefore exchange with solvent deuterium ($^2\text{H}_2\text{O}$) at a faster rate than backbone amide protons in less solvent accessible ordered regions. The degree of exchange over various time intervals is assessed by quenching the exchange kinetics by lowering the pH and temperature, fragmenting the protein by pepsin proteolysis, and measuring the mass of the resulting fragments by mass spectrometry. Briefly, a 5 μL aliquot of purified protein (25-50 μg) is mixed with 15 μL of $^2\text{H}_2\text{O}$ and incubated at 0 $^\circ\text{C}$. At several time intervals, 30 μL of Quench Solution (1.0 guanidine hydrochloride, 0.5% formic acid, $\sim\text{pH}$ 2.5) is added to stop the exchange reaction and partially unfold the protein (increasing efficiency of pepsin cleavage). These samples are then immediately flash frozen at -80°C . Samples are then thawed on ice and subjected to limited proteolysis (pepsin column liquid chromatography). The resulting peptides are separated by reverse phase liquid chromatography using a C18 column and acetonitrile gradient. The eluate is analyzed using an electrospray-linear ion-trap mass spectrometer (ESI-MS) and the resulting peptides and their masses are identified. Peptides with greater mass (higher deuterium exchange) compared to the fully protonated control are identified. The results are depicted graphically as a heat map (Figure 6B); residues in peptides with the greatest amount of mass increase (disordered regions) are represented with red boxes. Residues in peptides showing little or no mass increase (ordered regions) are represented by blue boxes. This information allows design of alternative constructs (removing disordered regions) with improved crystallization properties (Pantazatos et al., 2004; Spraggon et al., 2004) and providing spectra more amenable for NMR assignment and structure determination (Sharma et al., 2009), as shown in Figure 6C.

6.3 Wheat germ cell free protein expression

The NESG has made considerable gains in the area of eukaryotic expression and protein production during Protein Structure Initiative-2. However, analysis of the expression and solubility data derived from the NESG pipeline during this time period reveals the eukaryotic targets are still recalcitrant in comparison to prokaryotic proteins. Overall, 42% of bacterial targets produce constructs with *ES* values greater than 11. By contrast, only 34% of the human targets produce constructs that meet our usability score criteria. Clearly the production of human and other eukaryotic proteins in *E. coli* remains challenging and alternative expression system are needed to allow the pursuit of high value human targets proven intractable in bacterial expression systems. Eukaryotic cell-free systems often permit successful production of proteins that undergo proteolysis or accumulate in inclusion bodies during bacterial expression (Vinarov et al., 2006). Although the yields produced by wheat germ cell-free systems can be challenging for crystallization studies, recent advances in NMR microprobe technologies allow for structural determination using relatively small protein quantities (< 100 micograms) (Aramini et al., 2007) that can be readily produced by wheat germ cell-free expression systems even when they cannot be made in *E. coli* cell expression systems.

In order to exploit this technology together with the “NESG Multiplex Vector Set”, we have modified the Promega TnT wheat germ cell-free vector to allow ligase independent cloning of the same PCR products used in our normal T7-based cloning pipeline. Specifically, we have modified the wheat germ cell-free expression vector pTSHQn from Promega to be compatible with our pET15_NESG vector and contain the identical N-terminal 6X-His tag and multicloning site sequences. This allows for the direct comparison of the same protein sequence expressed in both *E. coli* and wheat germ cell-free system, while also allowing a single PCR amplification product to be cloned into both vectors/expression systems. We have used this modified wheat germ cell-free system to study the efficacy of this approach with 66 non-secreted human targets that were problematic (both in expression and/or solubility levels) in the prokaryotic expression system (Zhao et al., 2010). Following wheat germ cell-free expression we have found that ~70% of the insolubly expressed (in pET15_NESG) proteins were solubly expressed in wheat germ cell-free system. Overall, 52% of the human target proteins (34 of the 66) were solubly expressed in the wheat germ cell-free system. Although the yields of purified protein produced in this wheat germ cell-free system are limited (100 - 500 microgram levels), NMR microprobe technology and selective labeling strategies such as SAIL (“Stereo Array Isotope Labeling”) (Aramini et al., 2007; Kainosho et al., 2006) enable wheat germ cell-free protein expression as a valuable tool for preparing samples for NMR studies.

6.4 Total gene synthesis and codon optimization

One possible reason for the lower success rates with human and eukaryotic proteins in producing expression constructs meeting our usability criteria may relate to codon usage or other translation-related factors. Numerous studies have been published linking the differences in frequency of codon usage in *E. coli* (and tRNA pool) and the effects on expression of recombinant proteins of human or other origin with differing codon usage patterns (Ikemura, 1985; Kanaya et al., 1999). Other studies have shown that codon usage is

not the critical factor, rather the stability of mRNA secondary structure near the ribosomal binding site is the overriding factor (Kudla et al., 2009), or a combination of both (Supek and Muc, 2010). Total gene synthesis allows for codons and secondary structure to be optimized to maximize translation (Tian et al., 2004). To explore this technology and its efficacy for high-throughput protein production, some human genes have been optimized and synthesized for expression in *E. coli* (Codon Devices). We have compared the expression and solubility (*ES*) score of the optimized constructs versus natural sequences, and natural sequences with various codon enhanced strains (Acton and Montelione, unpublished results). These studies indicate that codon optimization can be very effective in improving both expression and solubility of human proteins in *E. coli* expression hosts, providing a means to produce proteins that otherwise fail in bacterial expression systems.

7. *E. coli* Single Protein Production System for Isotopic Enrichment

Perdeuteration is an invaluable approach to the study of proteins by NMR. This is especially true for larger proteins (Gardner and Kay, 1998; Kay and Gardner, 1997; Rosen et al., 1996; Venters et al., 1995) and membrane proteins (Fernandez et al., 2004; Hiller et al., 2008). ^2H , ^{13}C , ^{15}N -enriched protein samples are also required for certain strategies utilizing rapid, fully automated analysis of small protein structures (Tang et al., 2010; Zheng et al., 2003). Although perdeuteration is a valuable tool, the preparation of deuterated samples can be costly; isotope costs for preparing ^2H , ^{13}C , ^{15}N -enriched samples range from \$1500 - \$3,000 per liter. *E. coli* expression, the preferred host for most protein production for NMR studies, is often inhibited by $^2\text{H}_2\text{O}$ and time consuming, laborious acclimation steps must be employed for effective growth in deuterium (Venters et al., 1995). Together, these issues have limited the use of perdeuteration in NMR studies. However, the use of the condensed single protein production method in *E. coli* (Suzuki et al., 2007; Suzuki et al., 2006; Suzuki et al., 2005; Schneider et al., 2010) alleviates many of these problems including the cost barrier.

Briefly, the condensed single protein production system utilizes MazF endoribonuclease overexpression (of the MazE/MazF addiction module) to induce bacteriostasis. The toxin cleaves mRNA at ACA triplets, inhibiting the translation of *E. coli* proteins causing the cells to enter into a quiescent state. Heterologous genes can be expressed at high levels under these conditions by the removal of ACA triplets in their coding sequence by total gene synthesis or site directed mutagenesis. Since the cells are in a state of bacteriostasis, they (i) are not negatively effected by media containing deuterium and (ii) can be condensed 40-fold or more while producing heterologous proteins for several days. The condensation provides costs savings by decreasing the volume of deuterated media used, and concomitantly the amounts of isotope labeled carbon and nitrogen sources and/or other precursors.

Recently, two modified condensed single protein production systems have been developed. Both are engineered to decrease or delay the expression of the heterologous gene before the addition of $^2\text{H}_2\text{O}$ containing media and hence produce recombinant protein with higher levels of perdeuteration. This is achieved in the condensed single protein production(tet) system (Schneider et al., 2009) using a dual induction system, IPTG for MazF and anhydrous tetracycline inducing the *tet O 1* operator for the heterologous protein in a

modified Cold-Shock Vector (Takara Biosciences). Alternatively, IPTG inducible MazF variants were produced that lack tryptophan or histidine residues (Vaiphei et al., 2010). Expression of these variants and heterologous genes (cloned into a Cold Shock Vector) in Trp or His auxotrophic strains does not allow the production of the recombinant protein until Trp or His are added to the media. Briefly, protein-coding regions for a target are produced lacking ACA triplets and cloned into an appropriate Cold-Shock Vector. *E. coli* strains harboring a plasmid with an IPTG inducible MazF gene are transformed with the expression construct. Transformants are grown in 1 l of minimal media until OD₆₀₀ reaches ~0.5 units, and cold-shocked at 15 °C. IPTG is added to induce MazF expression for 2-3 hours. Either before or after IPTG induction, the culture is centrifuged to pellet the cells, the supernatant is discarded and the cells are condensed (resuspended in 10 – 100-fold decreased volume) in deuterated media. Finally, heterologous protein expression is induced (anhydrotetracycline) or permitted (addition of labeled Trp or His) and incubation at 15 °C is performed for at least 16 hours.

8. Conclusions

The main goal of this monograph is to inform the NMR community of the tools and strategies used by the NESG for high-throughput protein production of protein samples for NMR studies. The current weekly capacity of the NESG protein sample production pipeline is roughly 100 new expression constructs and analytical scale expression screens, 96 mid-scale fermentations/purifications, 72-preparative scale fermentations, and purification of 30-40 proteins in the tens of milligrams scale. In the last five years this pipeline has produced over 21,000 expression constructs. Some 7,000 of these have *ES* values meeting our usability criteria, with nearly 4,000 successful protein purifications, resulting in over 750 Protein Data Bank depositions (in addition to more than 200 PSB depositions in the first-phase Protein Structure Initiative-1 project). The platform has provided samples allowing determination of NMR resonance assignments and three-dimensional structures for more than 400 proteins or protein domains. It is designed to work with commercially available yet relatively inexpensive equipment such as Qiagen BioRobot 8000 automation systems and plasticware, inexpensive platform shakers and ÄKTApur FPLC systems among others. This approach allows for (i) scalability, as higher throughput can be attained through the acquisition of additional equipment or personnel, and (ii) duplication in industrial, core facilities or small consortiums of structural biology labs. The tools outlined here including DisMeta, Primer Prim'er, and our detailed protocols for high-throughput cloning, expression and purification are freely available (<http://www-nmr.cabm.rutgers.edu/labdocuments/proteinprod/index.htm>) and will hopefully prove to be a valuable resource for the biological research community.

Acknowledgments

We thank Profs. S. Anderson, C. Arrowsmith, G. DeTitta, W. Hendrickson, J. Hunt, M. Gerstein, M. Inouye, M. Kennedy, J. Marcotrigiano, B. Rost, T. Szyperski and L. Tong, along with all the current and former members of the Rutgers Protein Production Team and all members of the NESG Consortium, for valuable advice in the development of the NESG Protein Sample Production Platform. This work was supported grants from the National Institute of General Medical Sciences Protein Structure Initiative U54-GM074958 (to GTM) and U54-GM094597 (to GTM).

References

- Acton TB, et al. Robotic cloning and Protein Production Platform of the Northeast Structural Genomics Consortium. *Methods Enzymol.* 2005; 394:210–43. [PubMed: 15808222]
- Acton TB, et al. An Alternative Method for Generating Genomic DNA PCR Template and Expansion of 'Reagent Genomes' in preparation. 2010
- Aramini JM, et al. Microgram-scale protein structure determination by NMR. *Nat Methods.* 2007; 4:491–3. [PubMed: 17496898]
- Aslanidis C, de Jong PJ. Ligation-independent cloning of PCR products (LIC-PCR). *Nucleic Acids Res.* 1990; 18:6069–74. [PubMed: 2235490]
- Bendtsen JD, et al. Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol.* 2004; 340:783–95. [PubMed: 15223320]
- Bertini I, et al. The annotation of full zinc proteomes. *J Biol Inorg Chem.* 2010; 15:1071–1078. [PubMed: 20443034]
- Bertone P, et al. SPINE: an integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics. *Nucleic Acids Res.* 2001; 29:2884–98. [PubMed: 11433035]
- Brodsky O, Cronin CN. Economical parallel protein expression screening and scale-up in *Escherichia coli*. *J Struct Funct Genomics.* 2006; 7:101–8. [PubMed: 17187226]
- Chen L, et al. TargetDB: a target registration database for structural genomics projects. *Bioinformatics.* 2004; 20:2860–2. [PubMed: 15130928]
- Chikayama E, et al. Mathematical model for empirically optimizing large scale production of soluble protein domains. *BMC Bioinformatics.* 2010; 11:113. [PubMed: 20193068]
- Crowe J, et al. 6xHis-Ni-NTA chromatography as a superior technique in recombinant protein expression/purification. *Methods Mol Biol.* 1994; 31:371–87. [PubMed: 7921034]
- Dean FB, et al. Comprehensive human genome amplification using multiple displacement amplification. *Proc Natl Acad Sci U S A.* 2002; 99:5261–6. [PubMed: 11959976]
- Dessailly BH, et al. PSI-2: structural genomics to cover protein domain family space. *Structure.* 2009; 17:869–81. [PubMed: 19523904]
- Dyson MR, et al. Production of soluble mammalian proteins in *Escherichia coli*: identification of protein features that correlate with successful expression. *BMC Biotechnol.* 2004; 4:32. [PubMed: 15598350]
- Englander SW. Hydrogen exchange and mass spectrometry: A historical perspective. *J Am Soc Mass Spectrom.* 2006; 17:1481–9. [PubMed: 16876429]
- Everett JK, et al. Primer Prim'r: A web based server for automated primer design. *Journal of Functional and Structural Genomics.* 2004; 5:13–21.
- Fernandez C, et al. NMR structure of the integral membrane protein OmpX. *J Mol Biol.* 2004; 336:1211–21. [PubMed: 15037080]
- Ferre-D'Amare AR, Burley SK. Use of dynamic light scattering to assess crystallizability of macromolecules and macromolecular assemblies. *Structure.* 1994; 2:357–9. [PubMed: 8081751]
- Ferre-D'Amare, AR.; Burley, SK. *Methods in Enzymology.* Vol. 276. Academic Press; New York: 1997. *Dynamic Light Scattering in Evaluating Crystallizability of Macromolecule*; p. 157-166.
- Gardner KH, Kay LE. The use of ²H, ¹³C, ¹⁵N multidimensional NMR to study the structure and dynamics of proteins. *Annu Rev Biophys Biomol Struct.* 1998; 27:357–406. [PubMed: 9646872]
- Gill SR, et al. Metagenomic analysis of the human distal gut microbiome. *Science.* 2006; 312:1355–9. [PubMed: 16741115]
- Goh CS, et al. Mining the structural genomics pipeline: identification of protein properties that affect high-throughput experimental analysis. *J Mol Biol.* 2004; 336:115–30. [PubMed: 14741208]
- Goh CS, et al. SPINE 2: a system for collaborative structural proteomics within a federated database framework. *Nucleic Acids Res.* 2003; 31:2833–8. [PubMed: 12771210]
- Graslund S, et al. Protein production and purification. *Nat Methods.* 2008a; 5:135–46. [PubMed: 18235434]

- Graslund S, et al. The use of systematic N- and C-terminal deletions to promote production and structural studies of recombinant proteins. *Protein Expr Purif.* 2008b; 58:210–21. [PubMed: 18171622]
- Hartley JL, et al. DNA cloning using in vitro site-specific recombination. *Genome Res.* 2000; 10:1788–95. [PubMed: 11076863]
- Haun RS, Moss J. Ligation-independent cloning of glutathione S-transferase fusion genes for expression in *Escherichia coli*. *Gene.* 1992; 112:37–43. [PubMed: 1339364]
- Hendrickson WA. Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. *Science.* 1991; 254:51–8. [PubMed: 1925561]
- Hiller M, et al. [2,3-(13)C]-labeling of aromatic residues--getting a head start in the magic-angle-spinning NMR assignment of membrane proteins. *J Am Chem Soc.* 2008; 130:408–9. [PubMed: 18092784]
- Huang YJ, et al. Targeting the human cancer pathway protein interaction network by structural genomics. *Mol Cell Proteomics.* 2008; 7:2048–60. [PubMed: 18487680]
- Ikemura T. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol.* 1985; 2:13–34. [PubMed: 3916708]
- Jansson M, et al. High-level production of uniformly ¹⁵N- and ¹³C-enriched fusion proteins in *Escherichia coli*. *Journal of Biomolecular NMR.* 1996; 7:131–141. [PubMed: 8616269]
- Kainosho M, et al. Optimal isotope labelling for NMR protein structure determinations. *Nature.* 2006; 440:52–7. [PubMed: 16511487]
- Kanaya S, et al. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene.* 1999; 238:143–55. [PubMed: 10570992]
- Kay LE, Gardner KH. Solution NMR spectroscopy beyond 25 kDa. *Curr Opin Struct Biol.* 1997; 7:722–31. [PubMed: 9345633]
- Klock HE, et al. Combining the polymerase incomplete primer extension method for cloning and mutagenesis with microscreening to accelerate structural genomics efforts. *Proteins.* 2008; 71:982–94. [PubMed: 18004753]
- Krogh A, et al. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol.* 2001; 305:567–80. [PubMed: 11152613]
- Kudla G, et al. Coding-sequence determinants of gene expression in *Escherichia coli*. *Science.* 2009; 324:255–8. [PubMed: 19359587]
- Kvist T, et al. Specific single-cell isolation and genomic amplification of uncultured microorganisms. *Appl Microbiol Biotechnol.* 2007; 74:926–35. [PubMed: 17109170]
- Lamesch P, et al. hORFeome v3.1: a resource of human open reading frames representing over 10,000 human genes. *Genomics.* 2007; 89:307–15. [PubMed: 17207965]
- Lasken RS. Single-cell genomic sequencing using Multiple Displacement Amplification. *Curr Opin Microbiol.* 2007; 10:510–6. [PubMed: 17923430]
- Lasken RS. Genomic DNA amplification by the multiple displacement amplification (MDA) method. *Biochem Soc Trans.* 2009; 37:450–3. [PubMed: 19290880]
- Liu J, et al. Automatic target selection for structural genomics on eukaryotes. *Proteins.* 2004; 56:188–200. [PubMed: 15211504]
- Liu J, et al. Novel leverage of structural genomics. *Nature Biotech.* 2007; 25:849–851.
- Lupas A, et al. Predicting coiled coils from protein sequences. *Science.* 1991; 252:1162–4. [PubMed: 2031185]
- Markley JL, et al. Recommendations for the presentation of NMR structures of proteins and nucleic acids--IUPAC-IUBMB-IUPAB Inter-Union Task Group on the standardization of data bases of protein and nucleic acid structures determined by NMR spectroscopy. *Eur J Biochem.* 1998; 256:1–15. [PubMed: 9746340]
- Marsischky G, LaBaer J. Many paths to many clones: a comparative look at high-throughput cloning methods. *Genome Res.* 2004; 14:2020–8. [PubMed: 15489321]
- McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. *Bioinformatics.* 2000; 16:404–405. [PubMed: 10869041]

- Nair R, et al. Structural genomics is the largest contributor of novel structural leverage. *J Struct Funct Genomics*. 2009; 10:181–91. [PubMed: 19194785]
- Nallamsetty S, et al. Efficient site-specific processing of fusion proteins by tobacco vein mottling virus protease in vivo and in vitro. *Protein Expr Purif*. 2004; 38:108–15. [PubMed: 15477088]
- Neri D, et al. Stereospecific nuclear magnetic resonance assignments of the methyl groups of valine and leucine in the DNA-binding domain of the 434 repressor by biosynthetically directed fractional ¹³C labeling. *Biochemistry*. 1989; 28:7510–6. [PubMed: 2692701]
- Netzer WJ, Hartl FU. Recombination of protein domains facilitated by co-translational folding in eukaryotes. *Nature*. 1997; 388:343–9. [PubMed: 9237751]
- Pantazatos D, et al. Rapid refinement of crystallographic protein construct definition employing enhanced hydrogen/deuterium exchange MS. *Proc Natl Acad Sci U S A*. 2004; 101:751–6. [PubMed: 14715906]
- Peti W, et al. Towards miniaturization of a structural genomics pipeline using micro-expression and microcoil NMR. *J Struct Funct Genomics*. 2005; 6:259–67. [PubMed: 16283429]
- Porath J. Immobilized metal ion affinity chromatography. *Protein Expr Purif*. 1992; 3:263–81. [PubMed: 1422221]
- Price WN 2nd, et al. Understanding the physical properties that control protein crystallization by analysis of large-scale experimental data. *Nat Biotechnol*. 2009; 27:51–7. [PubMed: 19079241]
- Punta M, et al. Structural genomics target selection for the New York consortium on membrane protein structure. *J Struct Funct Genomics*. 2009; 10:255–68. [PubMed: 19859826]
- Qin J, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*. 2010; 464:59–65. [PubMed: 20203603]
- Rosen MK, et al. Selective methyl group protonation of perdeuterated proteins. *J Mol Biol*. 1996; 263:627–36. [PubMed: 8947563]
- Rossi P, et al. A microscale protein NMR sample screening pipeline. *J Biomol NMR*. 2010; 46:11–22. [PubMed: 19915800]
- Rost B, et al. The PredictProtein server. *Nucleic Acids Res*. 2004; 32:W321–6. [PubMed: 15215403]
- Rual JF, et al. Human ORFeome version 1.1: a platform for reverse proteomics. *Genome Res*. 2004; 14:2128–35. [PubMed: 15489335]
- Schneider WM, et al. Independently inducible system of gene expression for condensed single protein production (cSPP) suitable for high efficiency isotope enrichment. *J Struct Funct Genomics*. 2009; 10:219–25. [PubMed: 19642019]
- Schneider WM, et al. Efficient condensed-phase production of perdeuterated soluble and membrane proteins. *J Struct Funct Genomics*. 2010; 11:143–154. [PubMed: 20333498]
- Sharma S, et al. Construct optimization for protein NMR structure analysis using amide hydrogen/deuterium exchange mass spectrometry. *Proteins*. 2009; 76:882–94. [PubMed: 19306341]
- Sheibani N. Prokaryotic gene fusion expression systems and their use in structural and functional studies of proteins. *Prep Biochem Biotechnol*. 1999; 29:77–90. [PubMed: 10069435]
- Shirano Y, Shibata D. 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*. 1990; 271:128–30. [PubMed: 2121525]
- Slabinski L, et al. XtalPred: a web server for prediction of protein crystallizability. *Bioinformatics*. 2007; 23:3403–5. [PubMed: 17921170]
- Snyder DA, et al. Comparisons of NMR spectral quality and success in crystallization demonstrate that NMR and X-ray crystallography are complementary methods for small protein structure determination. *J Am Chem Soc*. 2005; 127:16505–11. [PubMed: 16305237]
- Spraggon G, et al. On the use of DXMS to produce more crystallizable proteins: structures of the T. maritima proteins TM0160 and TM1171. *Protein Sci*. 2004; 13:3187–99. [PubMed: 15557262]
- Studier FW, Moffatt BA. Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. *J Mol Biol*. 1986; 189:113–30. [PubMed: 3537305]
- Supek F, Muc T. On relevance of codon usage to expression of synthetic and natural genes in *Escherichia coli*. *Genetics*. 2010; 185:1129–34. [PubMed: 20421604]

- Suzuki M, et al. Single protein production (SPP) system in *Escherichia coli*. *Nat Protoc.* 2007; 2:1802–10. [PubMed: 17641648]
- Suzuki M, et al. Bacterial bioreactors for high yield production of recombinant protein. *J Biol Chem.* 2006; 281:37559–65. [PubMed: 17020876]
- Suzuki M, et al. Single protein production in living cells facilitated by an mRNA interferase. *Mol Cell.* 2005; 18:253–61. [PubMed: 15837428]
- Tang Y, et al. Fully automated high-quality NMR structure determination of small (²H)-enriched proteins. *J Struct Funct Genomics.* 2010 DOI: 10.1007/s10969-010-9095-6.
- Tian J, et al. Accurate multiplex gene synthesis from programmable DNA microchips. *Nature.* 2004; 432:1050–4. [PubMed: 15616567]
- Vaiphei ST, et al. Use of amino acids as inducers for high-level protein expression in the single-protein production system. *Appl Environ Microbiol.* 2010; 76:6063–8. [PubMed: 20656860]
- Venters RA, et al. High-level ²H/¹³C/¹⁵N labeling of proteins for NMR studies. *J Biomol NMR.* 1995; 5:339–44. [PubMed: 7647552]
- Vinarov DA, et al. Wheat germ cell-free platform for eukaryotic protein production. *FEBS J.* 2006; 273:4160–9. [PubMed: 16930128]
- Woods VL Jr. Hamuro Y. High resolution, high-throughput amide deuterium exchange-mass spectrometry (DXMS) determination of protein binding site structure and dynamics: utility in pharmaceutical design. *J Cell Biochem Suppl.* 2001; (Suppl 37):89–98. [PubMed: 11842433]
- Wootton JC, Federhen S. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* 1996; 266:554–71. [PubMed: 8743706]
- Xiao R, et al. The high-throughput protein sample production platform of the Northeast Structural Genomics Consortium. *J Struct Biol.* 2010; 172:21–33. [PubMed: 20688167]
- Zhao L, et al. Engineering of a wheat germ expression system to provide compatibility with a high throughput pET-based cloning platform. *J. Struct. Funct. Genomics.* 2010 DOI 10.1007/s10969-010-9093-8.
- Zheng D, et al. Automated protein fold determination using a minimal NMR constraint strategy. *Protein Sci.* 2003; 12:1232–46. [PubMed: 12761394]
- Zhu B, et al. In-fusion assembly: seamless engineering of multidomain fusion proteins, modular vectors, and mutations. *Biotechniques.* 2007; 43:354–9. [PubMed: 17907578]

A. Disorder Prediction

B. Alternative Constructs

Figure 1. The NESG Disorder Prediction Server (DisMeta) (<http://www-nmr.cabm.rutgers.edu/bioinformatics/disorder/>) and alternative constructs produced by the Construct Optimization Software for the *Porphyromonas gingivalis* protein Q7MX54 (NESG ID: PgR37)

(A) Output of the DisMeta server including prediction of secondary structure (PROFsec, PSIPred), parallel coiled coil regions (COIL) (Lupas et al., 1991), signal peptides (SignalP), transmembrane helices (TMHMM), low complexity regions (SEG) (Wootton and Federhen, 1996), and the Disorder Consensus plot showing the number of disorder prediction algorithms (0-8) predicting disorder versus the protein residue number in linear order N- to C- terminus. (B) Schematic representation of construct optimization for NESG protein target PgR37 from PFAM domain family DUF477, including: full-length, residues 54-187, residues 59-182 and residues 35-182. Only the 35-182 construct produced soluble expressed protein, and ultimately an NMR structure (Protein Data Bank ID:2KW7).

Figure 2. Schematic of the PLIMS directed high-throughput cloning process using the Qiagen BioRobot 8000

The PLIMS Database, utilizing a 96-well graphical user interface (GUI), generates DNA sequences corresponding to the protein regions selected by the NESG Construct Optimization Software. Primer Prim'er generates primer sets in 96 well format, each with the necessary 15 base pair overhangs for Infusion ligation-independent cloning. 96 well PCR reactions are performed, amplification products are visualized and separated by agarose gel electrophoresis. The results are archived in PLIMS and correct bands excised and purified using a custom 96 well-gel extraction protocol on the BioRobot. Purified amplification products contain 15-base pair regions of homology with each end of the linearized vector (shown as gray (5') and black (3') extensions on the PCR insert and vector in the lower center of the figure). Addition of the Infusion enzyme for pairing and resecting the region of overlap is followed by *E. coli* transformation. Colony PCR/agarose gel electrophoresis identifies the correct clones, the data is entered into PLIMS, which rearrays the colony PCR template plates for inoculation of 48 well blocks and an automated 96 well miniprep protocol on the BioRobot.

Figure 3. High-throughput analytical scale protein expression screening using robotic methods
A 96-well transformation protocol is performed on the BioRobot. Colonies are transferred to an LB containing 2.2 mL 96-well S-Block (Qiagen), followed by overnight subculturing in MJ9 media (all manipulations performed on the BioRobot deck, shown in lower left). Plates are agitated in a microtiter plate freezing rack attached to a platform shaker (shown in the lower middle). A 1:20 dilution into four 24 well blocks (Qiagen) is performed and cells are grown to mid-log phase and induced overnight at 17 °C. Following overnight incubation cells are harvested by centrifugation (3000 g-force, 10 minutes) and resuspended in 100 µL of Lysis Buffer and transferred to a 96-well Round Bottom plate (Greiner). Following sonication (Qsonix 96 probe with cell lysate shown in upper right) a 30 µL aliquot of the total cellular lysate (Tot) is transferred to a new plate. The remainder is centrifuged for 10 minutes at 3000 g-force, and a 30 µL aliquot of the supernatant (Sol) is transferred to a new plate. SDS-PAGE analysis of equal amounts of Total cell extract (left) or Soluble cell extract (right): lane 1, SDS-PAGE standard (Precision Plus, Bio-Rad), lanes 2 and 3 NESG target HR6654C (residues 207- 289, 11 kDa), lanes 4 and 5 NESG target HR203 (residues 18-247, 26 kDa), lane 6 NESG target HR6430A (residues 287-370, 10.4 kDa). The constructs of all three human proteins produce soluble overexpressed protein of the correct size (highlighted with an asterisk).

Figure 5. Preparative-scale fermentation

(A) Flow chart of the preparative-scale fermentation process, The PLIMS database identifies targets suitable for fermentation and their expression glycerol stock location. The precultures and initial log-phase growth are carried out at 37 °C, overnight induction is performed at 17 °C. (B) SDS-PAGE analysis of a fermentation with four human protein targets gel showing Total Expressed (E - lanes 2, 4, 6, and 8) and Soluble (S – lanes 3, 5, 7, 9) extracts. The intensity of the targeted protein band in the total extract is used to estimate the total expression level (E), and the intensity of the band in the S extract is used to estimate the portion of the expressed protein that is soluble (S).

Figure 6. Construct optimization using amide hydrogen deuterium exchange with mass spectrometry detection (HDX-MS)

(A) Schematic of the HDX-MS process. Proteins are mixed with $^2\text{H}_2\text{O}$, exchange is quenched at three time points (e.g., 10 sec, 100 sec and 1000 sec) by lowering the pH and temperature. The samples are then subjected to a pepsin column for digestion, followed by reverse phase chromatography for peptide separation, and the degree of amide exchange prior to quenching is assessed by ESI-MS. (B) Results of the HDX-MS analysis for *E. coli* yiaD (NESG target ID: ER553). Primary sequence is shown at the top followed by PROFsec secondary structure prediction and heat map of the HDX-MS results at 10, 100 and 1000 sec reaction intervals. Note that the numbering does not include the N-terminal lipoprotein signal sequence that is cleaved from the mature protein. (C) ^1H - ^{15}N HSQC spectra of full length (left) and construct optimized (right) *E. coli* yiaD (59-199).

Table I**Buffer Optimization for NMR Studies**

Buffer ID	pH	Buffer Formula
MJ001	6.5	20 mM MES [†] , 100 mM NaCl, 5 mM CaCl ₂ , 10 mM DTT [‡] , 0.02% NaN ₃ , Complete [#] , 10% D ₂ O
MJ002	5.5	20 mM NH ₄ OAc, 100 mM NaCl, 5 mM CaCl ₂ , 10mM DTT [‡] , 0.02% NaN ₃ , Complete [#] , 10% D ₂ O
MJ003	4.5	20mM NH ₄ OAc, 100mM NaCl, 5mM CaCl ₂ , 10mM DTT [‡] , 0.02% NaN ₃ , Complete [#] , 10% D ₂ O
MJ004	5	50 mM NH ₄ OAc, 10 mM DTT [‡] , 50 mM Arginine, 0.02% NaN ₃ , Complete [#] , 10% D ₂ O
MJ005	5	50 mM NH ₄ OAc, 10 mM DTT [‡] , 5% CH ₃ CN, 0.02% NaN ₃ , Complete [#] , 10% D ₂ O
MJ006	6	50 mM MES [†] , 10 mM DTT [‡] , 50 mM Arginine, 0.02% NaN ₃ , Complete [#] , 10% D ₂ O
MJ007	6	50 mM MES [†] , 10 mM DTT [‡] , 5% CH ₃ CN, 0.02% NaN ₃ , Complete [#] , 10% D ₂ O
MJ008	6.5	25 mM Na ₂ PO ₄ , 450 mM NaCl, 10 mM DTT [‡] , 20 mM ZnSO ₄ , Complete [#] , 0.02% NaN ₃ , 10% D ₂ O
MJ009	6.5	20 mM MES, 100 mM NaCl, 5% CH ₃ CN, 10mM DTT [‡] , 0.02% NaN ₃ , Complete [#] , 10% D ₂ O
MJ010	6.5	20 mM MES, 100 mM NaCl, 50 mM Arginine, 10mM DTT [‡] , 0.02% NaN ₃ , Complete [#] , 10% D ₂ O
MJ011	6.5	20 mM MES, 100 mM NaCl, 1% Zwitter [§] , 10 mM DTT [‡] , 0.02% NaN ₃ , Complete [#] , 10% D ₂ O
MJ012	6.5	20 mM MES, 100 mM NaCl, 50 mM ZnSO ₄ , 10 mM DTT [‡] , 0.02% NaN ₃ Complete [#] , 10% D ₂ O

[‡]DTT: DL-Dithiothreitol (Sigma),

[†]MES: 2-(N-morpholino)ethanesulfonic acid (Sigma),

[§]Zwitter: ZWITTERGENTS® 3-12 (CALBIOCHEM),

[#]Complete: Complete Protease Inhibitor Cocktail (ROCHE).