Published in final edited form as:

J Struct Funct Genomics. 2015 June ; 16(2): 67–80. doi:10.1007/s10969-015-9198-1.

Expression Platforms for Producing Eukaryotic Proteins: A Comparison of *E. coli* Cell-Based and Wheat Germ Cell-Free Synthesis, Affinity and Solubility Tags, and Cloning Strategies

David J. Aceti,

The Center for Eukaryotic Structural Genomics, Department of Biochemistry, University of Wisconsin at Madison, 433 Babcock Dr., Madison, Wisconsin 53706

Mitochondrial Protein Partnership, Department of Biochemistry, University of Wisconsin at Madison, 433 Babcock Dr., Madison, Wisconsin 53706

Transmembrane Protein Center, Department of Biochemistry, University of Wisconsin at Madison, 433 Babcock Dr., Madison, Wisconsin 53706

Craig A. Bingman,

The Center for Eukaryotic Structural Genomics, Department of Biochemistry, University of Wisconsin at Madison, 433 Babcock Dr., Madison, Wisconsin 53706

Natural Product Biosynthesis, BioSciences at Rice and Department of Chemistry, Rice University, Houston, TX 77251

Mitochondrial Protein Partnership, Department of Biochemistry, University of Wisconsin at Madison, 433 Babcock Dr., Madison, Wisconsin 53706

Transmembrane Protein Center, Department of Biochemistry, University of Wisconsin at Madison, 433 Babcock Dr., Madison, Wisconsin 53706

Russell L. Wrobel,

The Center for Eukaryotic Structural Genomics, Department of Biochemistry, University of Wisconsin at Madison, 433 Babcock Dr., Madison, Wisconsin 53706

Mitochondrial Protein Partnership, Department of Biochemistry, University of Wisconsin at Madison, 433 Babcock Dr., Madison, Wisconsin 53706

Transmembrane Protein Center, Department of Biochemistry, University of Wisconsin at Madison, 433 Babcock Dr., Madison, Wisconsin 53706

Ronnie O. Frederick,

The Center for Eukaryotic Structural Genomics, Department of Biochemistry, University of Wisconsin at Madison, 433 Babcock Dr., Madison, Wisconsin 53706

Mitochondrial Protein Partnership, Department of Biochemistry, University of Wisconsin at Madison, 433 Babcock Dr., Madison, Wisconsin 53706

Transmembrane Protein Center, Department of Biochemistry, University of Wisconsin at Madison, 433 Babcock Dr., Madison, Wisconsin 53706

Shin-ichi Makino,

The Center for Eukaryotic Structural Genomics, Department of Biochemistry, University of Wisconsin at Madison, 433 Babcock Dr., Madison, Wisconsin 53706

Mitochondrial Protein Partnership, Department of Biochemistry, University of Wisconsin at Madison, 433 Babcock Dr., Madison, Wisconsin 53706

Transmembrane Protein Center, Department of Biochemistry, University of Wisconsin at Madison, 433 Babcock Dr., Madison, Wisconsin 53706

Karl W. Nichols,

The Center for Eukaryotic Structural Genomics, Department of Biochemistry, University of Wisconsin at Madison, 433 Babcock Dr., Madison, Wisconsin 53706

Sarata C. Sahu,

The Center for Eukaryotic Structural Genomics, Nuclear Magnetic Resonance Facility at Madison, Department of Biochemistry, University of Wisconsin at Madison, 433 Babcock Dr., Madison, Wisconsin 53706

Nuclear Magnetic Resonance Facility at Madison, Department of Biochemistry, University of Wisconsin at Madison, 433 Babcock Dr., Madison, Wisconsin 53706

Lai F. Bergeman,

The Center for Eukaryotic Structural Genomics, Department of Biochemistry, University of Wisconsin at Madison, 433 Babcock Dr., Madison, Wisconsin 53706

Paul G. Blommel,

The Center for Eukaryotic Structural Genomics, Department of Biochemistry, University of Wisconsin at Madison, 433 Babcock Dr., Madison, Wisconsin 53706

Claudia C. Cornilescu,

The Center for Eukaryotic Structural Genomics, Nuclear Magnetic Resonance Facility at Madison, Department of Biochemistry, University of Wisconsin at Madison, 433 Babcock Dr., Madison, Wisconsin 53706

Nuclear Magnetic Resonance Facility at Madison, Department of Biochemistry, University of Wisconsin at Madison, 433 Babcock Dr., Madison, Wisconsin 53706

Katarzyna A. Gromek,

The Center for Eukaryotic Structural Genomics, Department of Biochemistry, University of Wisconsin at Madison, 433 Babcock Dr., Madison, Wisconsin 53706

Mitochondrial Protein Partnership, Department of Biochemistry, University of Wisconsin at Madison, 433 Babcock Dr., Madison, Wisconsin 53706

Transmembrane Protein Center, Department of Biochemistry, University of Wisconsin at Madison, 433 Babcock Dr., Madison, Wisconsin 53706

Kory D. Seder,

The Center for Eukaryotic Structural Genomics, Department of Biochemistry, University of Wisconsin at Madison, 433 Babcock Dr., Madison, Wisconsin 53706

Soyoon Hwang,

The Center for Eukaryotic Structural Genomics, Department of Biochemistry, University of Wisconsin at Madison, 433 Babcock Dr., Madison, Wisconsin 53706

John G. Primm,

The Center for Eukaryotic Structural Genomics, Department of Biochemistry, University of Wisconsin at Madison, 433 Babcock Dr., Madison, Wisconsin 53706

Mitochondrial Protein Partnership, Department of Biochemistry, University of Wisconsin at Madison, 433 Babcock Dr., Madison, Wisconsin 53706

Transmembrane Protein Center, Department of Biochemistry, University of Wisconsin at Madison, 433 Babcock Dr., Madison, Wisconsin 53706

Grzegorz Sabat,

The Center for Eukaryotic Structural Genomics, Department of Biochemistry, University of Wisconsin at Madison, 433 Babcock Dr., Madison, Wisconsin 53706

Frank C. Vojtik,

The Center for Eukaryotic Structural Genomics, Department of Biochemistry, University of Wisconsin at Madison, 433 Babcock Dr., Madison, Wisconsin 53706

Brian F. Volkman,

The Center for Eukaryotic Structural Genomics, Department of Biochemistry, Medical College of Wisconsin, 8701 W. Watertown Plank Rd., Milwaukee, WI 53226

Zsolt Zolnai,

The Center for Eukaryotic Structural Genomics, Department of Biochemistry, University of Wisconsin at Madison, 433 Babcock Dr., Madison, Wisconsin 53706

George N. Phillips Jr.,

The Center for Eukaryotic Structural Genomics, Department of Biochemistry, University of Wisconsin at Madison, 433 Babcock Dr., Madison, Wisconsin 53706

Natural Product Biosynthesis, BioSciences at Rice and Department of Chemistry, Rice University, Houston, TX 77251

John L. Markley, and

The Center for Eukaryotic Structural Genomics, Nuclear Magnetic Resonance Facility at Madison, Department of Biochemistry, University of Wisconsin at Madison, 433 Babcock Dr., Madison, Wisconsin 53706

Nuclear Magnetic Resonance Facility at Madison, Department of Biochemistry, University of Wisconsin at Madison, 433 Babcock Dr., Madison, Wisconsin 53706

Mitochondrial Protein Partnership, Department of Biochemistry, University of Wisconsin at Madison, 433 Babcock Dr., Madison, Wisconsin 53706

Brian G. Fox

The Center for Eukaryotic Structural Genomics, Department of Biochemistry, University of Wisconsin at Madison, 433 Babcock Dr., Madison, Wisconsin 53706

Transmembrane Protein Center, Department of Biochemistry, University of Wisconsin at Madison, 433 Babcock Dr., Madison, Wisconsin 53706

Abstract

Vectors designed for protein production in *Escherichia coli* and by wheat germ cell-free translation were tested using 21 well-characterized eukaryotic proteins chosen to serve as controls within the context of a structural genomics pipeline. The controls were carried through cloning, small-scale expression trials, large-scale growth or synthesis, and purification. Successfully purified proteins were also subjected to either crystallization trials or ^1H - ^{15}N HSQC NMR analyses. Experiments evaluated: (1) the relative efficacy of restriction/ligation and recombinational cloning systems; (2) the value of maltose-binding protein (MBP) as a solubility enhancement tag; (3) the consequences of in vivo proteolysis of the MBP fusion as an alternative to post-purification proteolysis; (4) the effect of the level of LacI repressor on the yields of protein obtained from *E. coli* using autoinduction; (5) the consequences of removing the His tag from proteins produced by the cell-free system; and (6) the comparative performance of *E. coli* cells or wheat germ cell-free translation. Optimal promoter/repressor and fusion tag configurations for each expression system are discussed.

Keywords

Structural genomics; NIH Protein Structure Initiative; wheat germ cell-free translation; maltose binding protein solubility tag; Flexi Vector cloning; Gateway cloning

Introduction

The production of purified protein is a limiting step in many research projects [1, 2]. Consequently, the choices of expression host and vector system are important variables contributing to the success of research in structural genomics and functional proteomics. Broadly applicable and validated protein production platforms are also valuable for many other types of research.

During the Protein Structure Initiative, CESG initially developed *Escherichia coli* (*E. coli*) expression vectors that were based on Gateway cloning [3] and included either T7 or T5 promoters regulated by the *lac* operator and *lacI^q* repressor. Autoinduction was used to produce target proteins as fusions with a dual tag, His(6 or 8) and MBP, for IMAC purification and promotion of solubility, respectively [4-7]; the tags were removed using TEV protease [8].

For wheat germ cell-free translation, CESG initially used the pEU expression vector [9]. Target genes were cloned by a standard restriction/ligation approach and an SP6 promoter was used for in vitro transcription of target mRNA [10]. The transcripts encode a viral 5' and 3' untranslated regions that enhance translation in the absence of typical eukaryotic 5'

capping and 3' polyadenylation and also fuse a non-cleavable His6 tag to the N-terminus of the target protein.

As a subsequent stage of vector development, CESG tested a unified cloning approach based on the Flexi Vector cloning system that allowed production of nucleotide sequence-verified expression vectors for both *E. coli* cell-based expression and wheat germ cell-free translation from a single PCR reaction [11, 12]. The Flexi Vector cloning system was demonstrated to give efficiencies of single-pass cloning and serial transfer to other expression vectors that were comparable to those given by the Gateway system [3, 12, 13]. This vector design also yielded target proteins that had the same amino terminal sequences after tag cleavage using TEV protease, which allowed better testing of the differences in target protein behavior derived from the expression and purification methods as opposed to variations in the target sequence.

In this work, we examined total expression, solubility, susceptibility to TEV protease, yield of purified protein, and either crystal formation or spectral dispersion in ^1H - ^{15}N HSQC as benchmarks for the performance of a complete structural genomics pipeline. The study was carried out using a set of 21 eukaryotic control proteins whose behavior was previously established from operation of the CESG structural genomics pipeline. Since the studies were carried out using both *E. coli* cells and wheat germ cell-free translation, an expanded comparison of these two expression systems was also achieved. Additional factors such as the role of LacI in the control of bacterial expression during autoinduction, the ability of maltose binding protein to promote stability of fused proteins, and the timing, efficiency, and downstream consequences of TEV protease processing of fusion proteins were also examined. The cumulative results give insight into interdependencies in vector design that contribute to success in production of eukaryotic proteins for structural genomics research.

Materials and Methods

Production pipeline summary

For proteins made in *E. coli* cells, the production pipeline used in this work consisted of cloning, small-scale expression testing, large-scale production of cell paste, purification, and screening for amenability for structure determination using either crystallization or ^1H - ^{15}N HSQC spectroscopy. An analogous production pipeline was used for wheat germ cell-free translation, consisting of cloning, small-scale expression and purification testing, large-scale protein translation, purification, and structure screening using ^1H - ^{15}H HSQC NMR spectroscopy. Details of the methods used, including the scoring criteria used to assign successful outcomes, are described below.

Target Genes

The properties of the 21 genes used as controls in this study are given in Table 1. The sources of the original clones are given in Online Resource 1.

Cloning

The *E. coli* expression vectors pVP16, pVP56K, pVP65K, pVP68K and pVP91A were constructed from pQE80 (Qiagen) using standard molecular cloning techniques [14]. The wheat germ cell-free translation vector pEUHis was a gift from Professor Yaeta Endo (Ehime University, Matsuyama, Japan) and was adapted for Flexi Vector cloning [9, 11, 12]. Circular maps and functional properties of the expression vectors are shown in Figure 1 and Table 2, respectively. Methods for primer design, PCR amplification, cloning with Gateway [3], Flexi Vector [11, 12] and restriction digestion [15], and sequence verification of cloned genes are published. A successfully cloned gene encoded the exact amino acid sequence desired, either by exact nucleotide match to the published gene sequence or by the presence of one or more silent mutations.

Small-scale expression testing

The *E. coli* strain B834-pRARE2, which is useful in SeMet labeling for phase determination in X-ray crystallography, was the expression host [16]. Small-scale trials evaluating expression, solubility, and tag cleavage were carried out using autoinduction in both chemically defined and complex medium [5]. Expression of fusion proteins was scored on a high/medium/weak scale [5] based on comparison of the stained intensity of target protein bands with bovine serum albumin added in known amounts to Coomassie-stained SDS-PAGE gels. Expression was given a high rating when greater than 8 µg of protein target was present in the standard electrophoresis experiment (representing 93 µl of 25 hr culture), a medium rating when the amount of target protein was between 2 µg and 8 µg, and a weak rating when less than 2 µg of target protein was estimated. Solubility and TEV protease cleavage efficiency were assessed in a similar manner. Due to ambiguities in identifying the in vivo-cleaved target bands, all pVP65 targets were advanced to large-scale growth.

WEPRO2240H wheat germ extract (Cell-Free Sciences, Yokohama, Japan) was used for cell-free translation. Small-scale expression testing using the bilayer method, the scoring of expression, and purification of His tagged proteins were done as described [17]. A high rating was assigned if greater than 2.5 µg of fusion protein was purified from a 25 µL reaction, a medium rating for 1.25 to 2.5 µg, and a weak rating for less than 1.25 µg.

Large-scale growth and protein purification

Except as noted, targets with either high or medium ratings for expression, solubility, and TEV protease cleavage were considered to be suitable for scale-up efforts. After large-scale growth in the SeMet-labeling medium, samples of the cell paste were analyzed as described for small-scale expression testing and work was stopped on any target with any weak ratings. Published methods for large-scale growth [5] and protein purification [18] were used for targets expressed in *E. coli*. For targets expressed using wheat germ cell-free translation, large-scale protein synthesis [17] and purification [19] protocols were used with minor modifications, and generally correspond to steps used for purification of proteins from *E. coli* cells. Thus, for fusion proteins containing a TEV protease cleavage site, the purification was by IMAC capture and step-wise elution. The eluted protein was exchanged into “desalting” buffer and incubated with TEV protease at ambient temperature for 4 h. Fusion tags were separated from the target by subtractive IMAC using Ni-Sepharose high-

performance chromatography resin (GE Healthcare, Piscataway, NJ), exchanged into desalting buffer, and concentrated by Amicon Ultra-4 (Millipore, Billerica, MA). Tag-free target proteins were further purified by gel filtration using Superdex chromatography resin (GE Healthcare), and exchanged into a final buffer according to the pI of the protein [20]. For crystallization trials, proteins with pI <5 or in the range of 7-8 were placed in 5 mM MES, pH 6, containing 50 mM NaCl and 0.3 mM TCEP; proteins with pI in the range of 5-6 or 8-9 were placed in 5 mM HEPES, pH 7, containing 50 mM NaCl and 0.3 mM TCEP; and proteins with pI >9 or in the range of 6-7 were placed in 5 mM HEPES, pH 8, containing 50 mM NaCl and 0.3 mM TCEP. For NMR studies, proteins with pI in the range 6-8 were placed in 10 mM Na acetate, pH 5, containing 100 mM NaCl, 3 mM NaN₃ and 5 mM DTT; proteins with pI <6 or >8 were placed in 10 mM MOPS, pH 7, containing 100 mM NaCl, 3 mM NaN₃ and 5 mM DTT.

Protein characterization

The masses of purified ¹⁵N- or SeMet-labeled proteins were determined at the University of Wisconsin-Madison Biotechnology Center by electrospray ionization mass spectrometry using an Applied Biosystems 3200 Q Trap LC/MS/MS system and/or by matrix assisted laser desorption/ionization using an Applied Biosystems 4800 MALDI TOF/TOF. When needed, protein identification was verified by tryptic proteolysis, and molecular weight measurements on the generated peptides were made using an Agilent 1100 series nanoLC/MSD Trap SL. Peptides were identified using Mascot software (Matrix Science, London).

Crystallization screens

Crystallization screens were conducted with a uniform methodology for paired samples. Screening was conducted in high-throughput plates with the vapor-diffusion method using up to three distinct general screens: a local screen called UW-192 (Online Resource 2); IndexHT; and SaltRX (Hampton Research, Aliso Viejo, CA). The UW-192 screening was performed in a Corning 3775 192 condition crystallization plate (Corning, NY) using 50 nL of protein and reservoir solution, while the IndexHT and SaltRX screens were set in Innovaplate™ SD-2 plates (Innovadyne, Rohnert Park CA) using 100 nL of both protein and reservoir solutions. In all cases, screens were set using a Mosquito crystallization robot (TTP Labtech, Cambridge, MA) with a single protein concentration as received from the purification. In cases where the sample volume was insufficient to support the use of all three crystallization screens, UW-192 was applied preferentially, followed by IndexHT. Replicate crystallization screens were created to allow crystal growth and imaging at 4 °C and 20 °C. Progress was monitored using two CrystalFarm imaging systems (Bruker, Madison WI) set at the two target temperatures. Images were collected on a standard schedule, with photographs taken on days 1, 3, 7, 14, 28, 48 and 72. Each image was graded twice to assure consistency across paired samples.

NMR spectroscopy

The ¹⁵N-labeled proteins for NMR screening were produced either by *E. coli* cells or wheat germ cell-free translation. The concentration of protein in the NMR samples used for HSQC

measurements ranged between 0.05 to 1.0 mM. All NMR samples were placed in 5 mm Shigemi tubes, and ^1H - ^{15}N HSQC NMR spectra were recorded at the National Magnetic Resonance Facility at Madison on a Varian Inova 600 spectrometer (Agilent Technologies, Santa Clara, CA) equipped with a cryogenic probe. The temperature of the samples was held at 25°C. NMR spectra (^1H and ^1H - ^{15}N HSQC) were processed using the program NMRPipe [21]. Protein samples were categorized by the following criteria: (a) chemical shift dispersion of the peaks, particularly the ^1HN dispersion, greater than 3 ppm; (b) the peak count of resolved ^1H - ^{15}N peaks between 80% and 120% of the expected number; and (c) uniformity of peak intensities. Samples were scored HSQC+ if all three criteria (a–c) were satisfied, HSQC+/- if criterion (a) was met while either (b) or (c) were not, otherwise it was scored as HSQC-. Targets with favorable initial HSQC spectra were stored at room temperature for 2 weeks and reanalyzed to assess stability; all HSQC+ proteins described in this work were stable over this time period.

Other

Error bars shown in figures represent the standard error ($(np)(1-p)$), where n is the number of attempts and p = successes/attempts. Experimental data were managed using the “Sesame” laboratory information management system [22].

Materials and data access

The nucleotide sequences and vectors contained in the Control Workgroup are available from the PSI Materials Repository (<http://psimr.asu.edu/>) in four of the expression vectors central to this study; pVP16, pVP68K, pVP65K, and pEU-His-FV. The vectors comprising the Control Workgroup are available from the PSI Materials Repository as empty backbone vectors suitable for de novo cloning (pVP16, clone ID EvNO00306669; pVP68K, EvNO00306669; pEU-His, EvNO00306667; pVP65K, EvNO00306668; pVP91A, EvNO00306666; pEU-His-FV, EvNO00084283). Experimental results obtained with these proteins are available from the NIH Protein Structure Initiative Target Track Structural Biology Target Registration Database (<http://sbkb.org/tt/>) by search using the CESG ID numbers given in the tables.

Results

CESG Control Workgroup

The Control Workgroup is a set of 21 protein structure targets, 20 eukaryotic and 1 viral, whose behavior is known at all stages of the CESG protein production and structure determination process (Table 1). It includes targets that yielded structures deposited in the Protein Data Bank, some that were progressing satisfactorily in *E. coli* or in the wheat germ cell-free system but were discontinued due to success in the other platform, and others that failed at various points in the structure determination pipeline. The targets come from different organisms and have a wide range of physical characteristics including molecular weights and pI values; some targets contain metals or cofactors, and catalytic activity assays or other distinguishing biophysical properties are also available for many (Table 1, Online Resource 1). Of 20 targets previously attempted in the *E. coli* system, 13 proteins could be purified in high yield (tens or hundreds of mgs) and 7 in low yield (generally less than 10

mgs). Of 21 targets attempted in *E. coli* or wheat germ cell-free or both, the combined systems produced proteins that allowed solution of 17 structures; 5 by NMR, 11 by X-ray crystallography, and 1 by both methods. In the last case, both X-ray and NMR structures were solved for the *Arabidopsis* target At1g77540.1 (PDB 2EVN and 1XMT), both produced from *E. coli* cells. The X-ray structure of another protein, *Arabidopsis* agmatine iminohydrolase, was solved twice using protein produced by *E. coli* cells or by cell-free translation (1VKP and 3H7K).

Construction of expression vectors

Figure 1 shows circular maps of the vectors used in this study and Table 2 compares important features of the vectors. pVP16 was used to provide benchmark behavior against which subsequent vector modifications were tested in *E. coli* cell-based production; this vector features Gateway cloning, a strong lacI^q repressor, a His8-MBP tag cleaved from the target in vitro with TEV, and final protein products differing from native only by Ser replacement of the initial Met. The derivatives pVP56K (not shown), pVP68K, pVP65K, and pVP91A allowed tests of consequences on the outcome of pipeline operations of lacI repressor type, varying N-termini, Flexi Vector cloning, *in vivo* or “self-cleaving” tag removal, and the use of MBP as a solubility enhancement tag. For cell-free translation, pEU-His was used to provide benchmark behavior, and pEU-His-FV allowed testing of Flexi Vector cloning and the impact of His-tag removal.

An early derivative, pVP56K, varied from pVP16 in incorporating Flexi Vector cloning and a lacI instead of lacI^q repressor. Alteration of lacI^q to lacI in pVP56K and subsequent vectors was based on our finding that lacI is preferable for the auto-induction of recombinant protein expression in *E. coli* [23]. To minimize primer size, we used cloning method where the TEV protease site was included in the vector backbone, resulting in inclusion of the SgfI restriction site-encoded tripeptide AIA in the cleaved target protein. While initial testing showed that proteins with AIA at the N-terminus could be purified and successfully used for structure determination [24, 25], expanded pipeline studies suggested that this sequence was associated with a lower-than-expected percentage of crystallized targets; 3.3% of a random set of 697 targets with an N-terminus of Ser-target gave crystal structures from pVP16 while 1.4% of a random, non-overlapped set of 220 targets with N-terminal AIA-target gave crystals from pVP56K (data not shown). For this reason, pVP68K, pVP65K, pVP91A, and pEU-His-FV were designed to give the same mature target protein, with Ser as the N-terminal residue after tag removal using TEV protease. The decision to include Ser at the C-terminus of the TEV protease recognition site also corresponded to the high efficiency of proteolysis reported for the sequence ENLYFQ/S [26].

Production of SeMet-labeled proteins for X-ray crystallography in *E. coli*

The *E. coli* expression vectors pVP16 (Gateway cloning, in vitro proteolysis of the His8-MBP fusion, lacI^q), pVP68K (Flexi Vector cloning, in vitro proteolysis, lacI), and pVP65K (Flexi Vector cloning, in vivo proteolysis, lacI^q) were compared for the production of crystals from SeMet-labeled proteins. Starting with 21 protein structure targets and following normal CESG pipeline protocols, the number of successes were tracked for each vector at cloning, small-scale testing, large-scale protein production, purification, and

crystallization screening (Figure 2; see also Online Resource 3 for representative purification, crystallization, NMR, and mass spectrometry data, and Online Resource 4 for scoring and other results for each target in each vector attempted).

Flexi Vectors pVP68K and pVP65K immediately suffered attrition of 1 of the 21 targets at the cloning step. PmeI or SgfI restriction sites internal to target sequences preclude simple use of the Flexi Vector system, and these are present in 1-2% of genes from various eukaryotes [12]. The targets studied here were chosen from 5186 cloned at CESG without consideration as to whether these restriction sites were present, and one contained an internal PmeI restriction site and thus failed in the Flexi Vectors. Multiple cloning attempts were allowed for other targets as in normal pipeline operation, and all were successfully cloned.

Following standard procedures as detailed in Materials and Methods, targets in pVP16 and pVP68K with weak ratings for expression, solubility, or tag cleavage at either small-scale testing or large-scale growth were eliminated. It was not possible, however, to test pVP65K at small-scale as the in vivo-cleaved target proteins were undetectable against the background of *E. coli* proteins in SDS-PAGE gels, therefore all pVP65K targets were passed forward to large-scale production. Despite this (at least nominal) advantage, and large-scale production expression, solubility, and cleavage ratings comparable to the other vectors, pVP65K-expressed proteins performed relatively poorly in purification and were not carried forward to crystallization. In contrast, proteins expressed from pVP16- and pVP68K performed with statistical equivalence in large-scale production, purification, and crystallization.

Production of ^{15}N -labeled proteins for NMR in *E. coli* cells

The performance of the same three *E. coli* expression vectors were compared for production of ^{15}N -labeled proteins using 10 control proteins with molecular weights less than 25 kDa, and therefore amenable to structure determination by NMR (Figure 3). The pVP16 and pVP68K vectors performed similarly through purification. As before, pVP65K was not subjected to small-scale testing due to the difficulty of scoring and thus suffered no attrition at that step; nonetheless it lagged the other vectors somewhat in large-scale production, and significantly in protein purification. Of the 10 targets, 7 were purified from pVP16 (4 at 10 mg or more), 7 from pVP68K (all 7 at 10 mg or more), while only 3 were purified from pVP65K (1 at 10 mg or more). Favorable HSQC spectra were obtained from 5 proteins expressed from pVP16 and 3 from pVP68K; the difference is found in one cloning failure in pVP68K due an internal PmeI site that disallowed Flexi Vector cloning, and one target with poor solubility in large-scale expression. For the latter target (GO.2361), a crystallizable SeMet-labeled sample was obtained from the same vector, likely due to a marginally better solubility rating that allowed it to proceed in that pipeline.

Timing of fusion tag proteolysis

In typical use, an MBP-target fusion protein would be purified and then treated with TEV protease in order to remove the tag. The proteolysis reaction often requires an extended incubation, may not go to completion, and often requires TEV protease in substantial

amounts. To address these constraints, the vector pMCSG19 was produced by the Midwest Center for Structural Genomics that allowed in vivo “auto-cleavage” of fusion proteins by the action of a constitutively expressed TVMV protease [27]. This vector system was tested on a set of 132 proteins from *Salmonella typhimurium*, and the overall conclusion of the study was that use of pMCSG19 represented a useful salvage pathway for some types of targets.¹ A similar Flexi Vector construct produced by CESC, pVP65K, also gave promising early results, including support of the automated preparation of several human stem cell proteins and a high resolution crystal structure of GFP containing the problematic Flexi Vector AIA tag at the N-terminus [16].

Given the relatively small amount of information available on this approach, we undertook a comparative examination of control workgroup proteins expressed in pVP65K, in which TVMV in vivo cleavage of the MBP tag releases a target with Ser as the N-terminal residue followed by a His tag and TEV cleavage site. In the work discussed above, this vector was compared to pVP16 and pVP68K, both of which produce fusions that are subjected to TEV cleavage in vitro. Combining the results of 21 X-ray (SeMet-labeled) targets and 10 NMR (¹⁵N-labeled) targets, cumulative production from pVP65K lagged significantly both in numbers of proteins purified (35% of attempts) and yields obtained (13% at 10 mg or more); in contrast, 58% and 65% of attempted proteins were purified from pVP16 and pVP68K, respectively, with 42% and 58% at 10 mg or more. Importantly, a number of proteins that failed during the in vivo cleavage experiment were successful when prepared as the MBP fusion and proteolyzed after purification. These results support the role of MBP fusions in the stabilization of eukaryotic (and perhaps other difficult) proteins, both during translation and residence within the host cell and also during the purification process. Thus it appears that premature removal of the MBP tag may have a strongly deleterious effect on the viability of all but the most robust category of targets chosen for structural genomics research.

Contribution of MBP

A number of studies have suggested that when *E. coli* maltose binding protein is expressed as an N-terminal fusion with target proteins, the added ~45 kDa domain acts as a solubility-enhancing tag and perhaps a folding chaperone [6, 28, 29]. MBP itself is highly expressed, highly soluble and folds slowly, providing a rationale for its favorable activity as a fusion tag.

In addition to the study described above in which the timing of cleavage was varied, the efficacy of the MBP solubility tag was tested by comparing results for the control proteins expressed from either pVP68K or pVP16 (both TEV protease-cleavable His8-MBP tag fusions) with the same proteins expressed from pVP91A, which produces TEV protease-cleavable His8 tag fusions. Owing to the vector design, the exact same polypeptide sequences were obtained after treatment with TEV protease.

¹Three structures of proteins produced by use of pMCSG19 have been reported; none are targets investigated in the *S. typhimurium* study.

Twenty-one targets were tested in pVP91A in the SeMet-labeling pipeline. Only 4 successfully passed small-scale expression testing using the solubility standard that we normally applied to MBP-tagged proteins; it was then decided this standard might not be applicable to His-tagged proteins, and so a comparison was made without regard to small-scale expression testing or large-scale production ratings for any of the vectors. The goal of the study was purified, tag-cleaved, and mass spectrometry-verified target proteins.

The His-MBP fusion format, in both forms, resulted in significantly greater numbers of targets produced as well as targets yielding 1 mg or more and 10 mg or more (Table 3). Thus, for the same eukaryotic protein targets, the His8-MBP fusion was clearly superior to the His8 fusion in yield of protein obtained after proteolytic removal of the tag.

Comparison of pVP16 and pVP68K

The vectors pVP16 and pVP68K performed similarly in production of proteins that crystallized or gave favorable NMR HSQC spectra (Figures 2 and 3). However, a benchmark yield of 10 mg or more was reached more often by pVP68K, a potentially important advantage of that vector. While it is not clear that this result reaches the level of significance, it is reinforced by totaled fusion and cleaved target yields (Table 4). The relatively poor yields from pVP65K and pVP91A further support the superiority of the His-MBP vectors.

Outcomes at the level of individual targets expressed from pVP16 and pVP68K are shown in Table 5 and failure points are noted. Although it is difficult to identify significant differences in this relatively small pool of data, it is suggestive that pVP16 tended to fail at expression level (6 instances vs. 1 in pVP68K) and tag cleavage (4 vs. 0 in pVP68K) while pVP68K tended to fail at solubility (5 vs. 1 in pVP16) and crystallization/HSQC (7 vs. 3 in pVP16).

Production of ¹⁵N-labeled Proteins Using Wheat Germ Cell-free Translation

The performance of cell-free vectors pEU-His (restriction cloning, non-cleavable His tag) and pEU-His-FV (FlexiCloning, TEV-cleavable His tag) in the production of 20 targets for NMR (¹⁵N-labeling) was compared (Table 6). Here the target range was expanded beyond 25 kDa to explore the capabilities of cell-free translation to produce larger proteins; for example, incorporation of SAIL-labeled amino acids [30] could make these accessible to structure solution by NMR. Targets were carried as far as small-scale IMAC purification, where a “High” rating indicates suitability for scale-up to structural work in our pipeline, a “Medium” rating is marginal, and those with a “Weak” rating are not carried forward. Many proteins throughout the molecular weight range were successfully produced and the two vectors behaved equivalently for 13 targets, while pEU-His gave a better outcome for 6 targets and pEU-His-FV for 1. One cloning failure occurred in each vector, that in pEU-His due to an inability to PCR-amplify the target, and in pEU-His-FV due to an internal PmeI site. The 5 remaining failures in pEU-His-FV were attributable to 3 cases in which the pEU-His-FV products were found to bind less strongly to the IMAC media and were thus unpurifiable, and one each in expression and solubility. Since the only difference in proteins derived from pEU-His and pEU-His-FV is the N-terminal linker region

(MHHHHHAIAENLYFQS-target for pEU-His-FV vs. MGHHHHHHLE-target or MGHHHHHHLEDP-target for pEU-His, depending on restriction enzymes used), these results suggest that the combination of the AIA and TEV protease recognition site may not be completely benign. Notably, one group has found that the presence of the TEV recognition site correlated with decreased solubility for 9 proteins [31].

The 15 proteins successfully produced from pEU-His-FV in small-scale (High or Medium ratings) were carried forward to production and purification of larger quantities to obtain ^1H - ^{15}N HSQC spectra with or without the His6 tag, in parallel trials (Figure 4). Although equivalent numbers of tag-cleaved and tag-on proteins were purified, 6 cleaved targets gave favorable HSQC spectra compared to 3 uncleaved, supporting the utility of tag removal as part of a robust structure determination pipeline. This may at least partially counteract the apparent penalty of the AIA residues and TEV recognition site in pEU-His-FV.

Discussion

The Control Workgroup

Because of its composition, the Control Workgroup has utility in defining the efficacy of new methods, including downstream consequences on expression, stabilization, purification, tag proteolysis, functional assays, and structure determination. The Control Workgroup is also a useful resource for process validation, error checking, and staff training. In addition to the Control Workgroup, individual vectors for the expression of an additional ~4800 eukaryotic proteins in *E. coli* and ~200 in wheat germ cell-free translation have also been deposited into the NIH Protein Structure Initiative Materials Repository[32]. The majority of these are constructed using the expression vectors described in this work.

E. coli protein production pipeline

The *E. coli* Flexi Vector pVP68K was found to be equivalent to pVP16 upon consideration of the cloning efficiency, the range of proteins purified, and the numbers of crystals or favorable HSQC spectra obtained. Somewhat greater quantities of protein were obtained from pVP68K (Figs. 2 and 3, Table 4), likely due to the weaker lacI repressor in that vector and the compatibility of lacI with autoinduction, resulting in a more favorable time course and more complete use of the available carbon sources [23]. However, we cannot rule out the possibility of a contribution from the choice of antibiotic selection; unlike the kanamycin resistance encoded by pVP68K, pVP16-encoded ampicillin resistance results in degradation of the antibiotic and eventual loss of selection. We did not explore this possibility through the use of identical markers on the plasmids or by using the more stable carbenicillin in place of ampicillin.

Higher expression levels from pVP16 correlated with decreased protein solubility (Table 5), a commonly seen relationship in recombinant protein production. Ultimately, pVP68K is preferable for general usage in a high-throughput pipeline using autoinduction, since the range of suitable samples produced is equivalent to that from pVP16 but average yield is greater. Particular proteins require a lower expression rate for proper folding and solubility,

and pVP16 represents a useful salvage pathway for such proteins. Alternatively, pVP16 might be useful in a *lacI* format, a conversion that could be simply accomplished by the PCR approach previously described [23].

The pVP68K vector may also have an advantage in tag cleavage, succeeding at this step in 3 cases where pVP16 failed. The tag-target linker of the two vectors are quite different; LINGDGAGLEVLFG/GPAIAENLYFQ/S in pVP68K and NSSNNNNNNNNNNLIGIDENLYLTSLYKKAGSENLYFQ/S in pVP16, where **ENLYFQ/S** is the TEV recognition and cut site in both vectors and LEVLFG/GP is an HRV 3C (human rhinovirus 3C) protease recognition and cut site [33] in pVP68K used as a backup cleavage site; experimentation relating linker composition and TEV cleavage efficiency might well be a fruitful area for research.

The utility of maltose-binding protein (MBP) as a solubility tag was investigated through two lines of experimentation; (1) the consequences of early His-MBP tag cleavage in pVP65K and, (2) direct comparison with His tag vector pVP91A. The results confirm the benefits of the MBP tag and counter the common perception that MBP results in only a temporary solubilization of the target that is lost when the tag is removed.

Problematic consequences of “auto-cleaving” vectors

Interest in self-cleavage has arisen from unfavorable aspects of TEV protease treatment: extra steps needed in automated operation, incomplete proteolysis reaction, and the need for a large amount of protease characterized by relatively slow turnover [34-36]. However, the auto-cleaving vector pVP65K, which constitutively produces TVMV protease that cleaves the His-MBP tag *in vivo*, performed poorly. This was most evident at the protein purification stage, where lower yields of the final targets were recovered from pVP65K as compared to either pVP16 or pVP68K. Moreover, it was particularly troubling that proteins readily obtained and whose structures were determined by using the MBP fusion format catastrophically failed in auto-cleaving format. Thus it appears that most, if not all, of the advantages conferred by the MBP tag are lost in the auto-cleaving format. Conceivably, conversion of TVMV production from constitutive expression to controlled induction in late growth or stationary phase could address this problem. However, if MBP also serves as a stabilizing carrier during purification, auto-cleavage may be ultimately deleterious to the outcome for many targets.

Alternative methods to rapidly remove tags in a highly specific manner are desirable. Some of these, such as SUMO [37], and subtilisin [38], are now reported. Systematic testing of their methods can be achieved using the Control Workgroup and comparative studies described here.

Influence of *lacI* versus *lacI^q* on protein yield

In the ¹⁵N-labeling experiment described above, pVP68K significantly outperformed pVP16 with respect to yield of protein recovered. In general, yields of both MBP fusions and cleaved target proteins obtained from autoinduction were higher from vectors with the *lacI* format (pVP68K; Table 4) than from the vector with the *lacI^q* format (pVP16), which

produces about 20-fold more LacI repressor. These results corroborate the conclusions made in our earlier study, using GFP and luciferase as reporter genes, of the mechanism and consequences of vector design and medium composition on the outcome of autoinduction [23]. In summary, autoinduction can be strongly inhibited by the use of the *lacI^q* gene in the transcriptional control system, such that full execution of the autoinduction process is not possible under many metabolic conditions. Tight transcriptional control of basal expression is better achieved by use of a non-inducing minimal medium composition [23].

Wheat germ cell-free translation pipeline

It is reasonable that a eukaryotic expression system might have superior capacity to produce eukaryotic protein targets when compared to a prokaryotic expression system. Indeed, the cumulative results of this study of Control Workgroup targets and an earlier study of 96 *Arabidopsis* targets [34] show that the wheat germ cell-free system is capable of producing most of the proteins that can be made in *E. coli* cells plus an additional group not accessible with *E. coli*. The automated capacity of the cell-free system also allows quick and cost effective screening for the expression, solubility, and automated purification of many targets [39], including integral membrane proteins [40-42]. Conversely, the cell-free system often has a yield disadvantage compared to *E. coli* when more than a few milligrams of protein are required.

The Control Workgroup

These studies have given insight into optimal properties of expression vectors for use in structural biology pipelines. For *E. coli* expression, a *lacO/lacI* system performed best in the autoinduction format. Tight control of basal expression was provided by catabolite repression in aerobic medium containing glycerol, and the shift to growth on lactose/glycerol during autoinduction gave higher yield of purified proteins. The efficacy of the MBP tag fusion was also confirmed, including significant advantages of retaining the tag relative to “self-cleavage” formats. Moreover, the benefits of removing small linkers or tags such as AIA- and His6-ENLYFQ- were demonstrated by crystallization and HSQC NMR screening experiments. The original, well-optimized pEU vector used for cell-free translation [42] was shown to give the best results for wheat germ cell-free expression. However, this advantage was offset by the relative complexity of restriction cloning without positive selection and the inability to remove tags from the expressed proteins. The pEU-His-FV vector provides new utility for structural genomics pipelines by overcoming two disadvantages of the original vector; it allows compatible cloning by a single PCR reaction into the *E. coli* Flexi Vectors described here, and enables removal of N-terminal tags by TEV protease cleavage.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported by National Institutes of Health (NIH) Grants from the Institute for General Medical Sciences P50 GM64598, U54 GM074901, U01 GM094622, U54GM094584, U01 GM098248 and, in support of

the National Magnetic Resonance Facility at Madison, P41 GM103399. The authors thank other members of the Center for Eukaryotic Structural Genomics team including Emily Beebe, Seth Burgie, Darius Chow, Leigh Grundhoefer, Andrew Larkin, Yuko Matsubara, Xiaokang Pan, Donna Troestler, and Steven Wells for helpful discussions.

Abbreviations

CESG	Center for Eukaryotic Structural Genomics
Flexi Vector	Promega Flexi Vector cloning system
HSQC	heteronuclear single quantum correlation
IMAC	immobilized metal affinity chromatography
IPTG	isopropyl β -D-1-thiogalactopyranoside
MBP	maltose binding protein
SeMet	selenomethionine
TEV	tobacco etch virus
TVMV	tobacco vein mottling virus
HRV 3C	human rhinovirus 3C protease

References

- Kim Y, Babnigg G, Jedrzejczak R, Eschenfeldt WH, Li H, Maltseva N, Hatzos-Skintges C, Gu M, Makowska-Grzyska M, Wu R, An H, Chhor G, Joachimiak A. High-throughput protein purification and quality assessment for crystallization. *Methods*. 2011; 55(1):12–28. [PubMed: 21907284]
- Savitsky P, Bray J, Cooper CD, Marsden BD, Mahajan P, Burgess-Brown NA, Gileadi O. High-throughput production of human proteins for crystallization: the SGC experience. *J Struct Biol*. 2010; 172(1):3–13. [PubMed: 20541610]
- Thao S, Zhao Q, Kimball T, Steffen E, Blommel PG, Ritters M, Newman CS, Fox BG, Wrobel RL. Results from high-throughput DNA cloning of *Arabidopsis thaliana* target genes using site-specific recombination. *Journal of structural and functional genomics*. 2004; 5(4):267–276. [PubMed: 15750721]
- Routzahn KM, Waugh DS. Differential effects of supplementary affinity tags on the solubility of MBP fusion proteins. *Journal of structural and functional genomics*. 2002; 2(2):83–92. [PubMed: 12836665]
- Sreenath HK, Bingman CA, Buchan BW, Seder KD, Burns BT, Geetha HV, Jeon WB, Vojtik FC, Aceti DJ, Frederick RO, Phillips GN Jr. Fox BG. Protocols for production of selenomethionine-labeled proteins in 2-L polyethylene terephthalate bottles using auto-induction medium. *Protein expression and purification*. 2005; 40(2):256–267. [PubMed: 15766867]
- Hammarstrom M, Hellgren N, van Den Berg S, Berglund H, Hard T. Rapid screening for improved solubility of small human proteins produced as fusion proteins in *Escherichia coli*. *Protein Sci*. 2002; 11(2):313–321. [PubMed: 11790841]
- Tsunoda Y, Sakai N, Kikuchi K, Katoh S, Akagi K, Miura-Ohnuma J, Tashiro Y, Murata K, Shibuya N, Katoh E. Improving expression and solubility of rice proteins produced as fusion proteins in *Escherichia coli*. *Protein expression and purification*. 2005; 42(2):268–277. [PubMed: 15914031]
- Parks TD, Leuther KK, Howard ED, Johnston SA, Dougherty WG. Release of proteins and peptides from fusion proteins using a recombinant plant virus proteinase. *Analytical biochemistry*. 1994; 216(2):413–417. [PubMed: 8179197]

9. Sawasaki T, Hasegawa Y, Tsuchimochi M, Kasahara Y, Endo Y. Construction of an efficient expression vector for coupled transcription/translation in a wheat germ cell-free system. *Nucleic Acids Symp Ser.* 2000; (44):9–10. [PubMed: 12903243]
10. Vinarov DA, Markley JL. High-throughput automated platform for nuclear magnetic resonance-based structural proteomics. *Expert Rev Proteomics.* 2005; 2(1):49–55. [PubMed: 15966852]
11. Blommel PG, Martin PA, Seder KD, Wrobel RL, Fox BG. Flexi vector cloning. *Methods in molecular biology.* 2009; 498:55–73. [PubMed: 18988018]
12. Blommel PG, Martin PA, Wrobel RL, Steffen E, Fox BG. High efficiency single step production of expression plasmids from cDNA clones using the Flexi Vector cloning system. *Protein expression and purification.* 2006; 47(2):562–570. [PubMed: 16377204]
13. Walhout AJ, Temple GF, Brasch MA, Hartley JL, Lorson MA, van den Heuvel S, Vidal M. GATEWAY recombinational cloning: application to the cloning of large numbers of open reading frames or ORFeomes. *Methods in enzymology.* 2000; 328:575–592. [PubMed: 11075367]
14. Sambrook, J.; Russell, DW. *Molecular Cloning: A Laboratory Manual.* Cold Spring Harbor Laboratory Press; Cold Spring Harbor, NY: 2001.
15. Vinarov DA, Lytle BL, Peterson FC, Tyler EM, Volkman BF, Markley JL. Cell-free protein production and labeling protocol for NMR-based structural proteomics. *Nat Methods.* 2004; 1(2): 149–153. [PubMed: 15782178]
16. Frederick RO, Bergeman L, Blommel PG, Bailey LJ, McCoy JG, Song J, Meske L, Bingman CA, Ritters M, Dillon NA, Kunert J, Yoon JW, Lim A, Cassidy M, Bunge J, Aceti DJ, Primm JG, Markley JL, Phillips GN Jr. Fox BG. Small-scale, semi-automated purification of eukaryotic proteins for structure determination. *Journal of structural and functional genomics.* 2007; 8(4): 153–166. [PubMed: 17985212]
17. Makino S, Goren MA, Fox BG, Markley JL. Cell-free protein synthesis technology in NMR high-throughput structure determination. *Methods in molecular biology.* 2010; 607:127–147. [PubMed: 20204854]
18. Jeon WB, Aceti DJ, Bingman CA, Vojtik FC, Olson AC, Ellefson JM, McCombs JE, Sreenath HK, Blommel PG, Seder KD, Burns BT, Geetha HV, Harms AC, Sabat G, Sussman MR, Fox BG, Phillips GN Jr. High-throughput purification and quality assurance of Arabidopsis thaliana proteins for eukaryotic structural genomics. *Journal of structural and functional genomics.* 2005; 6(2-3):143–147. [PubMed: 16211511]
19. Vinarov DA, Newman CL, Tyler EM, Markley JL, Shahan MN. Wheat germ cell-free expression system for protein production. *Curr Protoc Protein Sci.* 2006 Chapter 5:Unit 5 18.
20. Tanford, C. *Physical Chemistry of Macromolecules.* Wiley; New York: 1961.
21. Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *Journal of biomolecular NMR.* 1995; 6(3):277–293. [PubMed: 8520220]
22. Zolnai Z, Lee PT, Li J, Chapman MR, Newman CS, Phillips GN Jr. Rayment I, Ulrich EL, Volkman BF, Markley JL. Project management system for structural and functional proteomics: Sesame. *Journal of structural and functional genomics.* 2003; 4(1):11–23. [PubMed: 12943363]
23. Blommel PG, Becker KJ, Duvnjak P, Fox BG. Enhanced bacterial protein expression during auto-induction obtained by alteration of lac repressor dosage and medium composition. *Biotechnol Prog.* 2007; 23(3):585–598. [PubMed: 17506520]
24. Bae E, Bitto E, Bingman CA, McCoy JG, Wesenberg GE, Phillips GN Jr. Crystal structure of an eIF4G-like protein from Danio rerio. *Proteins.* 2010; 78(7):1803–1806. [PubMed: 20229607]
25. Bitto E, Bingman CA, Kondrashov DA, McCoy JG, Bannen RM, Wesenberg GE, Phillips GN Jr. Structure and dynamics of gamma-SNAP: insight into flexibility of proteins from the SNAP family. *Proteins.* 2008; 70(1):93–104. [PubMed: 17634982]
26. Kapust RB, Tozser J, Copeland TD, Waugh DS. The P1' specificity of tobacco etch virus protease. *Biochemical and biophysical research communications.* 2002; 294(5):949–955. [PubMed: 12074568]
27. Donnelly MI, Zhou M, Millard CS, Clancy S, Stols L, Eschenfeldt WH, Collart FR, Joachimiak A. An expression vector tailored for large-scale, high-throughput purification of recombinant proteins. *Protein expression and purification.* 2006; 47(2):446–454. [PubMed: 16497515]

28. Sachdev D, Chirgwin JM. Solubility of proteins isolated from inclusion bodies is enhanced by fusion to maltose-binding protein or thioredoxin. *Protein expression and purification*. 1998; 12(1): 122–132. [PubMed: 9473466]
29. Kapust RB, Waugh DS. *Escherichia coli* maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused. *Protein Sci*. 1999; 8(8):1668–1674. [PubMed: 10452611]
30. Kainosho M, Torizawa T, Iwashita Y, Terauchi T, Mei Ono A, Guntert P. Optimal isotope labelling for NMR protein structure determinations. *Nature*. 2006; 440(7080):52–57. [PubMed: 16511487]
31. Kurz M, Cowieson NP, Robin G, Hume DA, Martin JL, Kobe B, Listwan P. Incorporating a TEV cleavage site reduces the solubility of nine recombinant mouse proteins. *Protein expression and purification*. 2006; 50(1):68–73. [PubMed: 16798010]
32. Protein Structure Initiative Materials Repository.
33. Cordingley MG, Register RB, Callahan PL, Garsky VM, Colonno RJ. Cleavage of small peptides in vitro by human rhinovirus 14 3C protease expressed in *Escherichia coli*. *Journal of virology*. 1989; 63(12):5037–5045. [PubMed: 2555540]
34. Tyler RC, Aceti DJ, Bingman CA, Cornilescu CC, Fox BG, Frederick RO, Jeon WB, Lee MS, Newman CS, Peterson FC, Phillips GN Jr, Shahan MN, Singh S, Song J, Sreenath HK, Tyler EM, Ulrich EL, Vinarov DA, Vojtik FC, Volkman BF, Wrobel RL, Zhao Q, Markley JL. Comparison of cell-based and cell-free protocols for producing target proteins from the *Arabidopsis thaliana* genome for structural studies. *Proteins*. 2005; 59(3):633–643. [PubMed: 15789406]
35. Blommel PG, Fox BG. A combined approach to improving large-scale production of tobacco etch virus protease. *Protein expression and purification*. 2007; 55(1):53–68. [PubMed: 17543538]
36. Waugh DS. An overview of enzymatic reagents for the removal of affinity tags. *Protein expression and purification*. 2011; 80(2):283–293. [PubMed: 21871965]
37. Panavas T, Sanders C, Butt TR. SUMO fusion technology for enhanced protein production in prokaryotic and eukaryotic expression systems. *Methods in molecular biology*. 2009; 497:303–317. [PubMed: 19107426]
38. Ruan B, Fisher KE, Alexander PA, Doroshko V, Bryan PN. Engineering subtilisin into a fluoride-triggered processing protease useful for one-step protein purification. *Biochemistry*. 2004; 43(46): 14539–14546. [PubMed: 15544324]
39. Vinarov DA, Loushin Newman CL, Markley JL. Wheat germ cell-free platform for eukaryotic protein production. *FEBS J*. 2006; 273(18):4160–4169. [PubMed: 16930128]
40. Jarecki BW, Makino S, Beebe ET, Fox BG, Chanda B. Function of Shaker potassium channels produced by cell-free translation upon injection into *Xenopus* oocytes. *Scientific reports*. 2013; 3:1040. [PubMed: 23301161]
41. Adams PD, Grosse-Kunstleve RW, Hung LW, Ioerger TR, McCoy AJ, Moriarty NW, Read RJ, Sacchettini JC, Sauter NK, Terwilliger TC. PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr D Biol Crystallogr*. 2002; 58(Pt 11): 1948–1954. [PubMed: 12393927]
42. Aly KA, Beebe ET, Chan CH, Goren MA, Sepulveda C, Makino SI, Fox BG, Forest KT. Cell-free production of integral membrane aspartic acid proteases reveals zinc-dependent methyltransferase activity of the *Pseudomonas aeruginosa* prepilin peptidase PilD. *MicrobiologyOpen*. 2012

Fig. 1.
Maps of expression vectors used in this study. pVP16, pVP65, pVP68K and pVP91A are used for *E. coli* expression. pEU-His and pEU-His-FV are used for wheat germ cell-free translation

Fig. 2. Performance of *Escherichia coli* vectors pVP16, pVP68K and pVP65K in an X-ray crystallography production pipeline. Twenty-one ORFs were tested in three vectors for production of protein suitable for structure determination. “NA” is Not Attempted

Fig. 3.

Performance of three *Escherichia coli* vectors in production of ^{15}N -labeled proteins. Ten proteins from the control workgroup with molecular weights suitable for NMR structure determination were tracked for progress along the production pipeline described in Materials and Methods. Purified proteins were subjected to ^1H - ^{15}N NMR HSQC analyses. "NA" is Not Attempted

Fig. 4.

Performance of pEU-His-FV in production of ^{15}N -labeled proteins by wheat germ cell-free translation. Targets with high or medium ratings in Small-Scale Screening, as described in Materials and Methods, were advanced to large-scale production, purification, and ^1H - ^{15}N HSQC screening

Table 1

Proteins contained in the Control Workgroup.

Previous results ^a							Cell-free translation	
#	CESG ID	Protein Name	MW (Da)	PI	Source Organism	<i>Escherichia coli</i>		
1	GO.34382	cytoplasmic dynein light chain	10990	6.8	<i>Mus musculus</i>	low yield	NMR 1Y40	
2	GO.6042	unknown protein At1g7540.1	11736	7.9	<i>Arabidopsis thaliana</i>	high yield, X-ray 2EVN; NMR 1XMT	not studied	
3	GO.14751	thioredoxin h1	12673	5.6	<i>Arabidopsis thaliana</i>	low yield, assay	NMR 1XFL	
4	GO.33810	zinc finger protein	13169	8.7	<i>Homo sapiens</i>	not studied	NMR 1ZR9	
5	GO.81370	unknown heme-binding protein	16474	5.1	<i>Cyanidioschyzon merolae</i>	high yield, no crystals, HSQC-, cofactor	soluble	
6	GO.2361	unknown protein At1g01470.1	16543	4.7	<i>Arabidopsis thaliana</i>	high yield, crystals, NMR 1X08	HSQC+	
7	GO.11624	thioredoxin-like protein	17947	4.9	<i>Arabidopsis thaliana</i>	high yield, NMR 1XOY	not expressed	
8	GO.91592	allene oxide cyclase variant 1	19522	5.4	<i>Arabidopsis thaliana</i>	high yield, no cleavage	not studied	
9	GO.91593	allene oxide cyclase variant 2	21232	5.7	<i>Arabidopsis thaliana</i>	high yield, X-ray 1Z8K	soluble	
10	GO.605	phosphatase	24537	7.6	<i>Arabidopsis thaliana</i>	high yield, X-ray 1XRI, metal, assay	not studied	
11	GO.91571	enhanced C3 green fluorescent protein	26748	5.7	<i>Aequorea victoria</i>	high yield, X-ray 2QU1, cofactor, PTM, assay	not studied	
12	GO.91591	TEV protease	26922	8.4	Tobacco etch virus	high yield, assay	not studied	
13	GO.80048	sarcosine dimethylglycine methyltransferase	33324	5.9	<i>Galleria sulphuraria</i>	high yield, X-ray 2O57, assay	not studied	
14	GO.37540	glyoxylate/hydroxypyruvate reductase	35668	7.1	<i>Homo sapiens</i>	low yield, X-ray 2H1S, assay	not studied	
15	GO.79368	aspartoacylase	35735	6.1	<i>Homo sapiens</i>	low yield, X-ray 2I3C, assay	soluble	
16	GO.70653	dimetal phosphatase	36645	5.4	<i>Danio rerio</i>	low yield, X-ray 2NXXF, metal, assay	not studied	
17	GO.7312	putative steroid sulfotransferase	37140	5.4	<i>Arabidopsis thaliana</i>	high yield, X-ray 1Q44	not studied	
18	GO.8210	12-oxophytodienoate-10,11-reductase	42691	7.7	<i>Arabidopsis thaliana</i>	high yield, many crystal forms, X-ray 1Q45, cofactor, assay	expressed	
19	GO.24674	agmatine iminohydrolase	43156	5.1	<i>Arabidopsis thaliana</i>	high yield, X-ray 1VKP, assay	high yield, X-ray 3H7K	
20	GO.34351	unknown protein	50256	9.1	<i>Mus musculus</i>	low yield, X-ray 2GNX, metal	not studied	
21	GO.74329	<i>Photinus</i> (firefly) luciferase	60844	6.4	<i>Photinus pyralis</i>	low yield, assay	not studied	

^aResults from previous work in either *E. coli* cell-based expression or wheat germ cell-free translation that led to inclusion in the control workgroup. "X-ray" or "NMR" followed by a PDB identifier indicate that the structure was solved by the method stated. PTM, assay, metal, and cofactor indicates the existence of a post-translational modification, an activity assay, a metal cofactor, or a non-metal cofactor, respectively.

Table 2

Expression vectors used in this study

Vector	Platform	Cloning	Antibiotic ^a	Promotor/ Repressor	Encoded Fusion Protein ^b	Tag Cleavage ^c
pVP16	<i>E. coli</i>	Gateway	Amp	T5/lacI ^q	His8-MBP-TEV-S-target	TEV
pVP56K	<i>E. coli</i>	Flexi Vector	Kan	T5/lacI	His8-MBP-TEV-AIA-target	TEV
pVP68K	<i>E. coli</i>	Flexi Vector	Kan	T5/lacI	His8-MBP-AIA-TEV-S-target	TEV
pVP65K	<i>E. coli</i>	Flexi Vector	Kan	T5/lacI	MBP-TVMV-His8-AIA-TEV-S-target	TVMV/TEV
pVP91A	<i>E. coli</i>	Flexi Vector	Amp	T5/lacI	His8-TEV-S-target	TEV
pEU-His	cell-free	Restriction	Amp	SP6	His6-X ₂₋₄ -target	None
pEU-His-FV	cell-free	Flexi Vector	Amp	SP6	His6-AIA-TEV-S-target	TEV

^a Selection marker. Amp, ampicillin; Kan: kanamycin.^b His6 or His8, N-terminal histidine tag; MBP, maltose binding protein; TVMV, tobacco vein motting virus protease recognition sequence ETVRFQ/S; TEV, tobacco etch virus protease recognition sequence ENLYFQ; AIA, protein sequence corresponding to the *Sgf1* nucleotide sequence used for Flexi Vector cloning; S, PCR-derived substitution of the N-terminal Met with Ser to promote TEV protease action; X₂₋₄, 2 to 4 non-native residues whose identity varies depending on restriction enzymes used for cloning.^c TEV protease is used for in vitro cleavage of fusion proteins; TVMV protease is used for in vivo cleavage of fusion proteins.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 3

Comparison of target yields from a His-tag vector *versus* His-MBP-tag vectors.

Vector	Tag Format	% Targets Produced	% Yielding 1 mg	% Yielding 10 mg
pVP91A	His	43	38	24
pVP16	His-MBP	81	81	48
pVP68K	His-MBP	71	71	57

Table 4Yields obtained from *E. coli* expression vectors

			Yield (mg) ^a			
			Fusion Protein		Cleaved Target Protein	
Vector	Promoter/Repressor	Tag	mean	median	mean	median
pVP16	T5/lacI ^q	MBP	168	143	24	9
pVP68K	T5/lacI	MBP	264	293	28	24
pVP65K	T5/lacI	TVMV/MBP	60	16	4	0
pVP91A	T5/lacI	His8	21	6	9	0

^aBased on 33 attempts (combining purification of 22 SeMet-labeled and 11 U-¹⁵N-labeled proteins) except for pVP91A, which was based on 22 purifications of SeMet-labeled proteins.

Table 5

Progress ratings for targets expressed in *E. coli* with SeMet-labeling for crystallization or ^{15}N for HSQC NMR analysis. Each target in each vector tested is given either a positive score indicating successful crystallization or a suitable HSQC, or else the step at which production failed^a.

#	ORF ID	MW (Da)	pI	Crystallization		^1H - ^{15}N HSQC	
				pVP16	pVP68K	pVP16	pVP68K
1	GO.34382	10990	6.8	+	+	+	+
2	GO.6042	11736	7.9	+	+	+	+
3	GO.14751	12673	5.6	Exp	+	+	+
4	GO.33810	13169	8.7	Pur	Exp	Exp	Pur
5	GO.81370	16474	5.1	Crys	Crys	HSQC	HSQC
6	GO.2361	16543	4.7	+	+	+	Sol
7	GO.11624	17947	4.9	+	Clo	+	Clo
8	GO.91592	19522	5.4	Cle	+	Cle	HSQC
9	GO.91593	21232	5.7	Exp	Crys	Exp	HSQC
10	GO.605	24537	7.6	Pur	+	HSQC	HSQC
11	GO.91571	26748	5.7	+	+		
12	GO.91591	26922	8.4	Cle	Pur		
13	GO.80048	33324	5.9	+	+		
14	GO.37540	35668	7.1	Exp	Sol		
15	GO.79368	35735	6.1	Cle	Sol		
16	GO.70653	36645	5.4	+	Pur		
17	GO.7312	37140	5.4	Exp	Crys		
18	GO.8210	42691	7.7	+	+		
19	GO.24674	43156	5.1	+	+		
20	GO.34351	50256	9.1	Sol	Sol		
21	GO.74329	60844	6.4	+	Sol		

^a Cle, poor tag cleavage; Clo, cloning failure; Crys, crystallization failure; Exp, poor expression; Sol, poor solubility; Pur, purification failure; HSQC, poor NMR spectrum.

Table 6

Small-scale purification results for control proteins produced from wheat germ cell-free vectors. Quantity purified is given as High, Medium, or Weak¹. Where a difference between vectors is seen, the reason for the weaker rating is given in parentheses.

#	ORF ID	MW	pEU-His	pEU-His-FV
1	GO.34382	10990	High ^a	High
2	GO.6042	11736	High	High
3	GO.14751	12673	High	High
4	GO.33810	13169	High	High
5	GO.81370	16474	High	Medium (Purification)
6	GO.2361	16543	Medium	Weak (Expression)
7	GO.11624	17947	Medium	----- (Cloning)
8	GO.91592	19522	High	High
9	GO.91593	21232	High	High
10	GO.605	24537	High	Weak (Solubility)
11	GO.91571	26748	High	High
12	GO.80048	33324	High	High
13	GO.37540	35668	Medium	Medium
14	GO.79368	35735	High	High
15	GO.70653	36645	----- (Cloning)	High
16	GO.7312	37140	Medium	Weak (Purification)
17	GO.8210	42691	Medium	Medium
18	GO.24674	43156	High	High
19	GO.34351	50256	Medium	Weak (Purification)
20	GO.74329	60844	Medium	Medium

^a“High” indicates IMAC purification of > 3 µg fusion protein per 25 µL reaction, “Medium” indicates 1-3 µg, “Weak” indicates < 1 µg.