The protein production pipeline at the Center for Eukaryotic Structural Genomics: Retrospective analysis and further development

Won Bae Jeon, Frank C. Vojtik, Andrew C. Olson, Jason M. Ellefson, Janet E. McCombs, Hassan K. Sreenath, Paul G. Blommel, Kory D. Seder, Brendan T. Burns, Craig A. Bingman, Craig S. Newman, Brian G. Fox, and George N. Phillips Jr.

University of Wisconsin-Madison, Department of Biochemistry, 433 Babcock Drive, Madison, Wisconsin, USA 53706-1549, http://www.uwstructuralgenomics.org

Abstract

An *E. coli* cell-based protein production pipeline has been established for high-throughput structural determination at the Center for Eukaryotic Structural Genomics. We have collected a rich set of data in our Sesame LIMS system, and data on successes and failures are being analyzed to guide our target selection and purification processes. In an effort to improve the efficiency of protocols and quality of target proteins, we have generated a new vector that incorporates a $(His)_6$-tag at the N-terminus of the expression construct. His-MBP fusion proteins with $(His)_6$-tag have increased affinity for the IMAC resin compared to those with a $(His)_6$-tag. A 60 mM increase in imidazole concentration is required for elution of $(His)_6$-tag. Consequently, more extensive washing can be used to remove non-specifically bound, contaminating proteins without elution of the fusion protein. This change allows the ÄKTA Purifier system to perform a fully automatic bump elution and desalting process instead of a gradient elution with gel electrophoresis analysis for pooling fractions. The automatic multi-step purification process shortens the first step of the purification protocol by ~6 h without sacrificing protein purity, thus providing one full workday of a standard workweek for additional target polishing by gel filtration or ionic chromatography. The new protocol should provide more efficiently produced and better quality protein samples for X-ray or NMR structure determination.

CESG protein production pipeline

Expression vectors and large-scale cell growth

We have designed a $(His)_n$-MBP fusion tag system (n = 6 or 8) to overcome the low solubility of recombinant eukaryotic proteins and to provide a generic Ni-IMAC purification strategy. The pVP13 and pVP16 expression vectors used for these studies were derived from pQE80 (Qiagen, Valencia, CA) to express an N-terminal fusion protein consisting of $(His)_n$-MBP and a linker region containing the TEV protease site contiguous with the second residue of the target protein [1,2]. Either *E. coli* Rosetta or B834 strains were used to produce unlabeled-, ^{15}N-, and $^{15}N/^{13}C$-, or SeMet-labeled proteins, respectively [3,4]. Cells were inoculated in a 2-liter polyethylene terephthalate bottle which contained 500 ml of Terrific Broth or auto-induction medium and incubated in a shaker at 250 rpm, 25°C for 22-24 h. Cells were harvested by centrifugation at 5000 x *g* for 20 min. For detailed information, see posters presented by Paul Blommel, Ronnie Frederick, and Hassan Sreenath.

CESG protein purification protocols

The overall philosophy of the protein purification pipeline is to automate the protocols as much as possible while preserving protein quality sufficient for structural studies. Protein purification processes are as follows:

Step 1: Cell lysis and preparation of the soluble fraction
Step 2: 1st IMAC capture of His-tagged fusion proteins
Step 3: Desalting of fusion proteins into TEV proteolysis buffer
Step 4: TEV proteolysis of fusion tags
Step 5: 2nd IMAC removal of tag and isolation of target proteins
Step 6: Desalting of targets
Step 7: Concentration of targets
Step 8: Drop-freezing of targets

Steps 2, 3, 5, and 6 were performed on the ÄKTA Purifier controlled by Unicorn 4.12 software. Detail description of purification processes was reported [5].

1st IMAC capture of His-tagged fusion proteins

Automated purification of His-tagged MBP fusion proteins performed with an ÄKTA Purifier platform, which allows precise control of direct sample application onto the Ni affinity column, washing-out of contaminating proteins, and fusion protein elution by applying linear gradient of imidazole concentration. **Figure 1** demonstrates the example of multi-step processes for Ni-IMAC purification of 6 different fusion proteins. Fusion proteins were eluted from the columns between 100 and 300 mM imidazole concentrations. The fractions that show purity greater than 90% on SDS-polyacrylamide gel were combined and desalted for TEV proteolysis. A total of 885 purification trials were performed with an average yield of 134, 168, and 134 mg of fusion proteins from *Arabidopsis*, mouse, and human genomes, respectively.

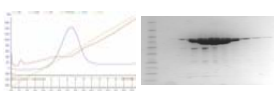

Figure 1. Chromatogram and SDS-polyacrylamide gel of the 1st IMAC capture of His-tagged MBP fusion protein from mouse genome, BC026994.

Cleavage of fusion tags by TEV protease

TEV protease is a cysteine protease [6]. Thus, it is crucial to keep the TEV protease under reducing conditions and to chelate metal ions which may catalyze the oxidation of the active site cysteine residue. After the 1st IMAC capture, EDTA (1 mM final concentration) was added to the combined fractions of fusion proteins to strip any Ni^{2+} ions leached from the Ni-IDA column. TEV protease reaction buffer also contains 0.3 mM TCEP and 100 mM NaCl which minimize the precipitation of fusion or target proteins during TEV proteolysis. In most cases, cleavage of His-MBP tag by TEV protease was complete after 3-5 h (**Fig. 2**). Poor cleavage of fusion proteins was occasionally observed. Uncleaved fusion proteins were analyzed by analytical gel filtration chromatography. Five of six uncleaved fusion proteins shown to form soluble protein aggregates (**Fig. 3A**) that presumably physically block the TEV protease cleavage site, since the monodispersed fractions were indeed cleavable (**Fig. 3B**).

Figure 2. Kinetics of His-tagged fusion protein cleavage by TEV protease. Reaction was performed at 17°C with a TEV protease to fusion protein ratio of 1 to 100. All proteins were from the mouse genome.

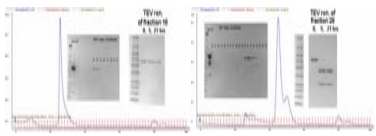

Figure 3. Analytical gel filtration chromatography was used to separate soluble aggregate from monodispersed forms of fusion protein. (**A**) At3g47470 (SeMet-labeled, uncleaved 63350 Da, cleaved 15483 Da). Fusion protein was eluted at void volume, fraction 15-19, indicating that it formed soluble aggregate. When TEV proteolysis was performed with the fraction 16, no cleavage was observed. (**B**) At2g31670 (SeMet-labeled, uncleaved 75136 Da, cleaved 27209 Da). Fusion protein was eluted at between 13.5 and 15 ml of eluent, fraction 27-30; indicating that protein is in monodispersed form. When TEV proteolysis was performed with the fraction 28, 95% of fusion protein was cleaved.

Isolation of target proteins

During the subtractive IMAC removal of His-tagged MBP and $(His)_6$-TEV protease, some target proteins eluted in the flow-through fractions (**Fig. 4A**), and some proteins bound weakly to the Ni-IDA column (**Fig. 4B**). The target proteins bound to the column might contain intrinsic, surface exposed His residues that form a metal complex with Ni^{2+} ions. Alternatively, target proteins may also be retained by hydrophobic and/or ionic interactions with the column matrix or with the cleaved His-MBP tag. Typically, the column-bound target proteins were eluted at the concentration of 70 mM imidazole. Chromatographic and SDS-PAGE analyses revealed that a two phase gradient elution was most effective to separate target proteins from His-MBP tags (**Fig. 4B**). Generally, the purity of most of the target proteins was greater than 90%. When the purity of target protein was less than 80%, polishing steps using either ion exchange and/or size exclusion improved the purity to greater than 90% (**Fig. 5**). Polishing programs were established with an ÄKTA HPLC system, and 49 proteins were improved in purity to greater than 90% from 123 attempts. The cumulative results for purified target proteins include 181 native, 96 SeMet-labeled, and 46 ^{15}N-labeled proteins with average yields of 43.3, 31.9, and 20.1 mg, respectively.

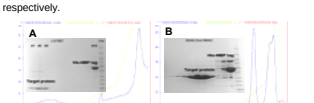

Figure 4. Chromatograms and SDS-polyacrylamide gels of the 2nd IMAC removal of His-MBP tags from target proteins. (**A**) Protein from the mouse genome, BC026994. (**B**) Protein from *Arabidopsis thaliana* genome, At2g22530.1.

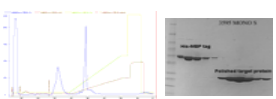

Figure 5. Chromatograms and SDS-polyacrylamide gels of target proteins polished by MonoS column chromatography. Protein from the *Arabidopsis thaliana* genome, At2g14110.

Overall performance

In PSI-1, a total of ~900 purifications were undertaken, with >450 giving purified protein and with >325 of these yielding more than 10 mg of target. Purification was judged successful if more than 3 mg of target protein with purity greater than 90% was drop-frozen for quality assurance. Overall purification success rate was about 50%.

Implementation of mass spectrometry for quality assurance

Of 315 proteins examined, 38 were found to be unsuitable for structural analysis (12%). Of those, nine were extensively degraded, 10 were cleanly truncated, 15 were incorrectly identified, one was poorly incorporated with SeMet, and three preparations had poor signal to noise in mass analysis due to low concentration. Incorporation of SeMet, ^{15}N, and ^{13}C isotopes was also determined by ESI-MS as described elsewhere [5]. Fourteen of 27 SeMet-labeled proteins exhibited incorporation of 90% or better; 10 more showed 80%-90% incorporation. Each of the 12 ^{15}N-labeled and two $^{15}N/^{13}C$-labeled proteins investigated were >95% labeled. Proteins not easily confirmed as correct by initial mass spectrometry were further investigated by proteolytic digestion and LC-MS/MS or amino acid analysis.

Retrospective analysis

Isoelectric point (pI) vs expression or solubility

Distribution of the calculated pI values of all targets cultured in the large-scale format was bimodal, with highest modes at pI approximately 5.5 and 9.0 for acid and basic proteins (**Fig. 6A**). Acidic proteins (pI<7) scored higher expression (**Fig. 6B**) and solubility (**Fig. 6C**) than basic proteins (pI>7.0), indicating that acidic targets are more expressed and soluble than basic targets in *E. coli* grown using current large-scale culture protocols.

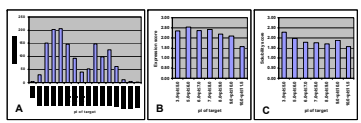

Figure 6. Distribution of the calculated pI values of targets (**A**). A total of 1322 large-scale (2-liter) cultures were grown with a set of 843, 259, 209 proteins from *Arabidopsis*, human, and mouse genomes, respectively. The pI of His-MBP tag is ~5.5. The pI of fusion protein is (pI of fusion - 5.5)/2. Average expression (**B**) and solubility (**C**) scores for each pI group. Expression and solubility were assessed according to CESG protocols. For quantitative analysis of experimental data, target proteins as scored high, medium, weak, or not expressed (insoluble) were converted to the number 3, 2, 1, or 0, respectively.

Molecular weight vs. purification success rate

Distribution of the MW of targets subjected to purification was unimodal, with an average MW of 26.8 kDa (**Fig. 7A**). The purification success rates of targets with MW values in the 10-50 kDa were 39-42%, while those with MW greater than 50 kDa showed a 23% of success rate (**Fig. 7B**). Targets with MW values less than 10 kDa showed the lowest success rate of purification (16%) (**Fig. 7B**). Thus our current pipeline works best for protein between 10-50 kDa, and other methods may be needed for smaller (<10 kDa) and larger (>50 kDa) proteins.

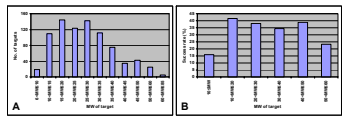

Figure 7. Distribution of the calculated MW values of targets (**A**). Average expression and solubility scores for each MW group (**B**).

Cysteine count in a target vs purification success rate

Distribution of the number of cysteine in a target was unimodal, with an average of 3.5 cysteines per target (**Fig. 8A**). For targets with cysteine count greater than 3, the purification success rate decreased as the numbers of cysteine in a target increased (**Fig. 8B**). Our target selection process discriminates against proteins with large number of cysteine, and the purification success data generally support this strategy.

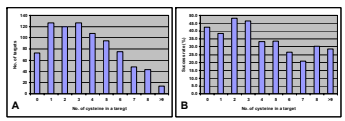

Figure 8. Distribution of the cysteine count in a target (A). Purification success rate for each cysteine count group (B).

Fraction of cysteine vs purification success rate

Distribution of the fraction of cysteine in a target was unimodal, with an average value of 1.60% (**Fig. 9A**). For targets with cysteine fraction values greater than 0.5%, the purification success rate decreased linearly (r^2 = 0.987) with increasing percentage of cysteine in a target (**Fig. 9B**).

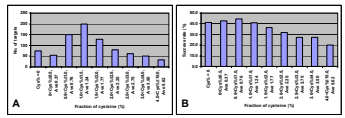

Figure 9. Distribution of the fraction of cysteine in a target (**A**). Purification success rate vs. fraction of cysteine (**B**).

Isoelectric point vs purification success rate

Distribution of the calculated pI values of targets from *Arabidopsis*, human, and mouse genomes was bimodal, with highest modes at pI approximately 5.5 and 9.0 for acid and basic proteins (**Fig. 10A**). The purification success rate decreased linearly (r^2 = 0.983) as pI values of target increased (**Fig. 10B**). The cause of this relationship is under investigation.

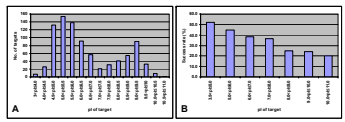

Figure 10. Distribution of the calculated pI values of targets (A). Purification success rate for each pI group (B).

Further development

Improvement of current pipeline protocols

Experimental bioinformatic data are used to improve efficiency of current *E. coli* protein production pipeline. The most common reason for failures of target purification is poor TEV proteolysis, particularly when the percentage of cleavage is less than 70% (**Table 1**). The results emphasize the requirement for screening methods that can assess the TEV proteolysis at small-scale cell culture stage. We are in the process of testing expression vectors that contain different proteases cleavage sites and different linker sequences and are developing high-throughput methods for screening for cleavage and solubility at the small-scale expression stage (see posters presented by Paul Blommel and Ronnie Frederick).

Small-scale purification

Trials (using PSI-1 targets successful in 2-liter growths) have shown that 8 mL of auto-induction culture is sufficient to yield >100 µg of purified protein with greater than 95% purity. In addition to the 250-fold reduction in reagents, the total process time for purification of 18 targets was ~24 h, as compared to ~96 h for our present large-scale purification efforts. When coupled with microfluidics (see poster presented by Craig Bingman), these improvements should result in a four-fold increase in the rate of producing samples for crystallization screening.

Table 1. Percentage of TEV proteolysis and purification success rate.

TEV proteolysis (%)	CESG category	No. of proteins subjected to purification	No. of proteins purified	Success rate (%)
⩽30	W	110	9	8.2
30<cleavage⩽40	M	47	1	2.1
40<cleavage⩽ 50	M	43	8	18.6
50<cleavage⩽60	M	31	6	19.4
60<cleavage⩽70	H	33	3	9.1
70<cleavage⩽80	H	36	13	36.1
⩾90	H	381	278	73.0

References:

1. Blommel, P.G., and Fox, B.G. (2005) *Anal. Biochem.*, 336, 75-86.
2. Thao, S., Zhao, Q., Kimball, T., Steffan, E., Newman, C.S., Fox, B.G., and Wrobel, R.L. (2004) *J. Struct. Funct. Genom.*, in press.
3. Tyler, R.C., Sreenath, H.K., Aceti, D.J., Bingman, C.A., Singh, S., Markley, J.L., and Fox, B.G. (2005) *Protein Expression Purif.*, in press.
4. Sreenath, H.K., Bingman, C.A., Buchan, B.W., Seder, K.D., Burns, B.T., Geetha, H.V., Jeon, W.B., Vojtik, F.C., Aceti, D.J., Frederick, R.O., Phillips, G.N. Jr., and Fox, B.G. (2005) *Protein Expression Purif.*, in press.
5. Jeon, W.B., Aceti, D.J., Bingman, C.A., Vojtik, F.C., Olson, A.C., Ellefson, J., McCombs, J.E., Sreenath, H.K., Blommel, P.G., Seder, K.D., Burns, B.T., Geetha, H.V., Harms, A.C., Sabat, G., Sussman, M.R., Fox, B.G., and Phillips, G.N., Jr. (2005) *J. Struc. Func. Genomics*, in press.
6. Phan, J., Zdanov, A., Evdokimov, A.G., Tropea, J.E., Peters, H.P.K., Kapust, R.B., Li, M., Wlodawer, A., and Waugh, D.S. (2002) *J. Biol. Chem.* **277**, 50564-50572.

CESG is supported by the National Institute of General Medical Sciences through the Protein Structure Initiative (P50 GM 64598).