Center for Eukaryotic Structural Genomics

Protein Structure Initiative

High-Throughput Purification of *Arabidopsis thaliana* Overexpressed in *E. coli* for Eukaryotic Structural Genomics

Won Bae Jeon, David Aceti, Craig Bingman, Hassan Sreenath, Frank Vojtik, Andrew Olson, Jason Ellefson, Janet McCombs, Paul Blommel, Kory Seder, Blake Buchan, Brendan Burns, Ronnie Frederick, John Kunert, Holalkere Geetha, Brian Fox, and George Phillips Jr.* *Corresponding Author

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

Abstract

An efficient and semi-automated pipeline system has been established for high-throughput purification of *E. coli*-expressed *A. thaliana* proteins. The key steps of this pipeline are (1) initial purification of $(His)_6$-MBP- tagged fusion proteins from cell lysates; (2) TEV protease cleavage; (3) removal of the liberated $(His)_6$-MBP tag from the target protein; and (4) target protein evaluation and concentration. The system currently handles 18 targets per week with a typical success rate of 42% (purification of 7 proteins to greater than 90% purity). TEV protease cleavage results in the addition of a single residue (Ser) to the N-terminus, replacing the native N-terminal methionine.

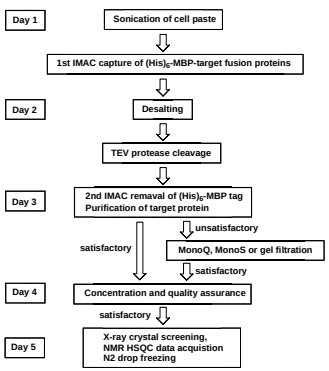
Optimized protocols for the purification of selenomethionine, ^{15}N, and $^{15}N;^{13}C$-labeled proteins are also available and commonly used, resulting in purity suitable for structural determination by X-ray crystallography and NMR. An on-line interactive laboratory information management system (the Lamp module of the Sesame software package [1]) is being used for data capture and analysis. Details and statistics for each purification step, including % TEV cleavage, yields, and purity, are presented in this poster.

All purified proteins are subjected to MALDI-TOF and ESI mass spectrometry to confirm identity and integrity, determine oligomeric state, and investigate possible ligands. Of 195 proteins examined, 143 passed inspection (139 monomers, 3 apparent dimers, and 1 apparent trimer). Nineteen samples failed (10 were truncations or degradations, 6 were targets mixed or misplaced during cloning, growth, or purification, 1 was poorly labeled with SeMet, 2 had uncharacterized mass anomalies). Incorporation of selenomethionine and ^{15}N and ^{13}C isotopes are also determined by ESI-MS. Fourteen of 27 selenomethionine-labeled proteins exhibited incorporation of 90% or better; 10 more showed 80%-90% incorporation. Each of 12 ^{15}N-labeled and 2 $^{15}N;^{13}C$-labeled proteins were >95% labeled. Proteins not easily resolved by initial mass spectrometry are further investigated by proteolytic digestion and LC-MS/MS or amino acid analysis. In addition, ICP-MS is performed to determine metal content; an initial round of ICP-MS analysis of 15 samples detected Mg^{2+} and Ca^{2+} ligands.

1. Z. Zolnai *et al.*, J. Structural and Functional Genomics **2**, 181-193 (2002).

Two-Step Purification of Recombinant Protein

The process employs a 2-step chromatography procedure, with an optional polishing step where ionic or size exclusion columns are used to improve the purity of the target protein. Once proteins have been processed through the concentration stages, all samples are sent to the ESI-MS and MALDI-MS analysis for quality assurance before X-ray crystal screening or NMR HSQC data acquisition. The current work flow begins with 18 cell pastes and typically provides 7 new proteins in a week (42% success rate).

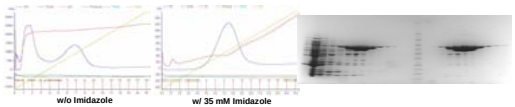


Optimization of Protein Purification Protocol

In order to optimize the pipeline, we developed protocols for automated protein production. We found that the presence of chaotrophic agents such as ethylene glycol and imidazole in the initial sonication buffer are critical to obtain high purity of fusion proteins. Optimum concentration of ethylene glycol and imidazole are 20% (w/v) and 35 mM, respectively. The optimized protocol allows purification of native, SeMet-, ^{15}N-, and $^{13}C;^{15}N$-labeled proteins up to 150 mg of protein from 2L culture volume. The purity of target proteins is typically greater than 90%, suitable for structural determination by X-ray crystallography and NMR.

Sample Elution Profile (ORF 15140, At3g12460, Hypothetical Protein) from 1st IMAC Capture

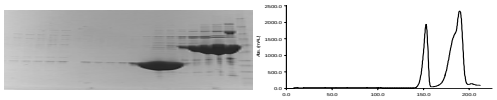
Method:
Sonication/Wash Buffer: 20 mM Na_2HPO_4, pH 7.5, 500 mM NaCl, ethylene glycol 20 %(w/v), 0.3 mM TCEP
IMAC-A Buffer: 20 mM Na_2HPO_4, pH 7.5, 500 mM NaCl, 0.3 mM TCEP
IMAC-B Buffer: 20 mM Na_2HPO_4, pH 7.5, 500 mM NaCl, 350 mM imidazole, 0.3 mM TCEP
Column: HiTrap™ Chelating HP 5 ml (Amersham Biosciences)
Elution: linear gradient from 10% IMAC-B to 80% IMAC-B over 150 ml



2nd IMAC Removal of $(His)_6$-MBP Tag from the TEV Protease Cleavage Mixture of Target Protein (ORF 8126, At2g33470, Unknown Protein)

During the 2nd IMAC removal of $(His)_6$-MBP tag, some target proteins were found in flow through fractions and some proteins were bound to Ni-column and eluted at the concentration of 50 mM imidazole.

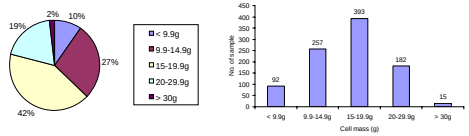
Method:
IMAC-A Buffer: 20 mM Na_2HPO_4, pH 7.5, 500 mM NaCl, 0.3 mM TCEP
IMAC-B Buffer: 20 mM Na_2HPO_4, pH 7.5, 500 mM NaCl, 350 mM imidazole, 0.3 mM TCEP
Column: HiTrapTM Chelating HP 5ml (Amersham Biosciences)
Elution: linear gradient from 100% IMAC-A to 30% IMAC-B over 50 ml and then to 100% IMAC-B over 45 ml, wash with 100% IMAC-B over 10 ml.



Statistics from the Protein Production

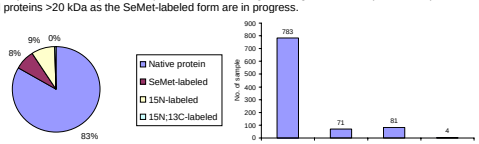
1. High-Throughput Cell Growth

CESG has adapted the 2L PET-bottle culture method (Millard, S., et al. 2003. *Protein Expr. Purif.* **29**:311-320) to provide high density cell culture in a cost-effective manner. The posters from CESG presented by Frederick et al. and Tyler et al. give further details on the approaches used for cell growth and isotopic labeling. We have performed large-scale growths on 939 samples. The average masses from 2L culture were 15.7 g, 17.7 g, and 16.7 for native, SeMet-, and ^{15}N-labeled protein, respectively. The chart below summarizes the distribution in cell mass obtained over these 939 growths and additional statistics.



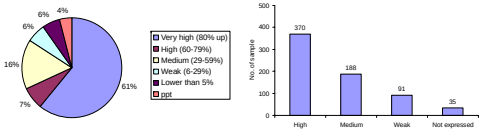
2. Production of Native, SeMet-, ^{15}N, and $^{15}N;^{13}C$-Labeled E. coli Cells

CESG developed labeling culture system using *E. coli* to produce SeMet-, ^{15}N, and $^{15}N;^{13}C$-incorporated protein in a high throughput mode (see CESG posters from Frederick, et al. and Tyler, et al.). To date, 783 native proteins (83%), 71 SeMet- (7.6%), 81 ^{15}N- (8.6%), and 4 $^{15}N;^{13}C$-labeled (0.4%) proteins have been produced. All proteins less than 20 kDa are now produced initially as ^{15}N-labeled proteins to permit NMR screening. Investigations on the potential to produce all proteins >20 kDa as the SeMet-labeled form are in progress.



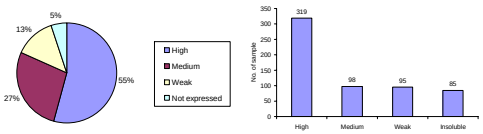
3. Overexpression of $(His)_6$-MBP-Tagged Fusion Proteins in E. coli

Advantages of the *Escherichia coli* expression system include low cost, ease of culture and high expression. A total of 684 samples that passed small scale expression evaluation were cultured on a 2L scale. SDS-PAGE analysis showed that 558 samples (82%) contained a high- or medium overexpressed fusion protein.



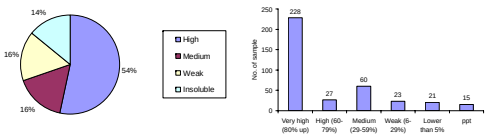
4. High Solubility of $(His)_6$-MBP-Tagged Fusion Proteins in E. coli

Some recombinant proteins form inclusion bodies under overexpression conditions. CESG has developed a (His)$_6$-MBP double tag system to overcome the low solubility of some recombinant proteins and to support a generic purification strategy. The MBP tag enabled the solubilization of ~86% of the fusion proteins examined, or 512 out of 597 tested samples.



5. Effective and Highly Specific Cleavage of Fusion Proteins

CESG produces TEV protease for internal use with greater than 95% purity. TEV protease is a cysteine protease. Thus, it is crucial to get rid of Ni^{2+} leached out from the immobilized Ni-column and to keep the TEV protease under reducing condition. During purification of target proteins, prior to desalting, EDTA (1 mM final concentration) was added to the 1st IMAC fractions that contain fusion proteins. TEV protease cleavage buffer also contains 0.3 mM TCEP and 100 mM NaCl which minimize the precipitation of fusion or target proteins during the TEV proteolysis. As shown below, ~68% of 374 samples showed >60% cleavage. Only 2 cases of non-specific internal cleavage have been detected and 15 samples precipitated upon TEV protease cleavage.

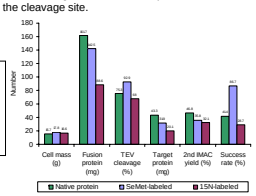


6. Comparison of the Yield, TEV Cleavage and Success Rate for the Purification of Native, SeMet-, and ^{15}N-Labeled Proteins

In order to identify issues in processing native, SeMet-, and ^{15}N-labeled proteins, the purification data were sorted and compared in terms of yield, TEV protease cleavage and success in passing the protein onward for structure determination efforts. TB, Studier, and minimal media provide similar levels cell mass production (column 1, 2, and 3). The amount of total fusion protein from *E. coli* grown on minimal medium was lower as compared to the cells grown on either TB or Studier media. The ^{15}N-labeled proteins were also less amenable to TEV protease cleavage (column 4, 5, and 6) resulting in lower yields of target protein (column 13, 14, 15) and overall success rate (column 16, 17, and 18). To increase the success rate of ^{15}N-labeled protein, the minimum amount of cell paste will be increased. In addition, CESG is evaluating new vectors to improve the 1st IMAC capture and the accessibility of TEV protease to the cleavage site.

Number of native protein examined was 357
Number of SeMet-labeled protein examined was 71
Number of ^{15}N-labeled protein examined was 81
Total number of native protein examined was 509

SeMet-proteins were selected on the basis of successful crystallization of native proteins and regrown using Studier medium, so exhibit a higher success rate.



Overview of Quality Assurance By Mass Spectrometry

Quality assurance of proteins by mass spectrometry was begun in May, 2003 and now covers all proteins produced by CESG, whether intended for X-ray crystallography or NMR, expressed in *E. coli* or wheat germ cell-free reactions, unlabeled or labeled with ^{15}N, $^{15}N;^{13}C$, or selenomethionine.

Of 181 proteins examined, 19 were found to be unsuitable for structural analysis (Table 1). Of 19 proteins rejected, 10 showed degradation, 6 were not the expected protein (due to pipeline errors), 1 had poor incorporationof selenomethionine label, and 2 had uncharacterized mass anomalies.

Table 1

STEP	ACCEPTED	REJECTED
1. Primary MALDI-TOF and ESI-MS	163	9
2. 2nd Round MALDI-TOF and ESI-MS	9	2
3. Proteolytic Digest and LC/MS-MS	7	8
4. Resequencing of Expression Vectors	2	0
TOTALS	181	19

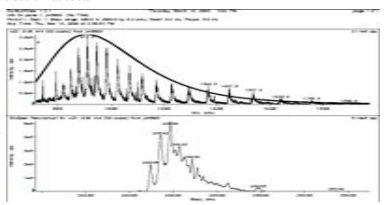
Determination of Selenomethionine Incorporation by Mass Spectrometry

Selenomethionine (SeMet) incorporation has been determined by ESI mass spectrometry for approximately 90% of the SeMet-labeled proteins produced by CESG. For the remaining 10%, the mass spectra are not clearly interpretable, and amino acid analysis followed by LC/MS/MS is being used.

To illustrate, the SeMet-labeled protein At5g01750.1, consisting of 217 residues and 7 methionines, was subjected to ESI-MS (Figure 1). The predicted molecular weight with full incorporation of SeMet is 24580 Da. Thus the 24586 Da peak represents the fully-labeled protein. Furthermore, the 24495 Da and 24542 Da mass peaks likely represent proteins with SeMet substituted for 5 and 6 methionine positions, respectively. Peaks with larger mass than the full incorporation peak are sodium adducts.

Figure 1

Data collected by Greg Sabat and Amy Harms, Mass Spectrometry/ Bioanalytical Facility, Biotechnology Center, University of Wisconsin-Madison.



Conclusion

An *Escherichia coli* expression system is being successfully used by CESG to overproduce *Arabidopsis thaliana* proteins for structural genomics. The advantages of this system includes cost-effectiveness, ease of culture, fast cell growth, and high expression. CESG has developed a $(His)_6$-MBP double tag system to overcome the low solubility of overexpressed recombinant proteins and to provide a generic purification strategy. The MBP-tag enables the solubilization of ~85% of 597 fusion proteins tested so far. With optimized protocols based on immobilized Ni-affinity chromatography, ~42% of 509 proteins have been purified with a yield and purity sufficient for structural studies by NMR and X-ray crystallography.

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