Vectors designed with inherent flexibility to support high-throughput expression screening and protein production

Paul G. Blommel, Russell L. Wrobel, Eric A. Steffan, Zachary Eggers, Won Bae Jeon,
Petar Duvjnak, Peter A. Martin, Ronnie O. Fredrick, Miriam F. Halstead, and Brian G. Fox
University of Wisconsin-Madison, Department of Biochemistry, 433 Babcock Drive, Madison, Wisconsin, USA 53706-1549, http://www.uwstructuralgenomics.org

Abstract

The Center for Eukaryotic Structural Genomics (CESG) has identified a strong correlation between positive results in screening for expression, solubility and protease cleavage and recovery of acceptable yields of purified protein from large-scale purification efforts. Furthermore, mining of the CESG databases also revealed targets that failed to meet one or more screening criteria but that nevertheless remained attractive for recycling into alternative expression and/or purification procedures. In order to effectively access multi-path evaluations of target protein behavior, an increased reliance on automation will also be required for process evaluations. To better support examination of these alternatives, our recent research has been directed toward expanding the options available for executing these recycling alternatives. In this poster, we describe a modular vector system with options for swapping solubility tags, expression promoters, protease specificities, antibiotic resistance, and cloning methods. Emphasis has been placed on vector designs and procedures that allow fluorescence-based quantification of protein expression, solubility, protease cleavage, and purification yield in formats compatible with high-throughput operations. Progress with the use of these constructs and the alternative uses inherent in their design are also illustrated.

Background

Prior to the implementation of a protein production pipeline, CESG investigated several options to optimize production of soluble eukaryotic proteins in *Escherichia coli*. This evaluation of affinity purification methods, expression systems, solubility tags, and cloning systems lead to the first generation pipeline production vector, pVP13-GW, whose map is shown in **Figure 1**.

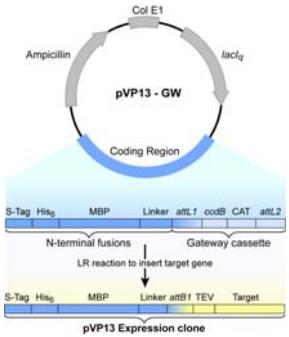

Figure 1. Plasmid map of pVP13-GW, the first generation vector developed at CESG for cell-based protein expression.

pVP13-GW was created from pQE80 (Qiagen, Valencia, CA) to contain an S-Tag for detection of protein levels, a His_6 tag as an affinity handle, and *E. coli* maltose binding protein (MBP) for solubility enhancement as a contiguous *N*-terminal fusion to the target protein. In addition, the Gateway™ recombination sequences were included to allow facile cloning of target proteins. Using the 2-step CESG cloning scheme, a TEV site is incorporated during amplification of the target gene, which allows release of the target protein from the *N*-terminal fusions with the minimal change of a serine substituted for the *N*-terminal methionine.

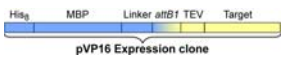

Figure 2. Plasmid map of pVP16, the current vector used for cell-based protein expression at CESG.

Figure 2 shows the critical region of the second generation *E. coli* production vector used at CESG. pVP16 was designed to provide tighter binding of fusion proteins to the immobilized metal affinity chromatography (IMAC) resin used for the primary purification of fusion proteins. This was accomplished by the addition of two additional histidine residues and by deletion of the S-Tag. Fusion proteins generated from pVP16 are eluted at higher imidazole concentrations than with pVP13-GW. Owing to the increased affinity, the pVP16 fusion proteins can be washed with higher concentrations of imidazole to more effectively remove contaminating *E. coli* proteins. Then, elution with a stepwise increase in imidazole rather than a gradient gives a more highly concentrated sample suitable for automated desalting (see poster by Won Bae Jeon for more information).

The X-Vector Concept

The next generation of CESG expression vectors was developed to address issues arising from continued operation of our protein production pipeline. One goal for the X-vector design was to increase the options available for expression of a soluble protein, including targets that failed for a variety of reasons during previous expression or purification attempts. In addition, we investigated ways to replace time-consuming SDS-PAGE screening for expression, solubility, and protease processing with instrumental measurements compatible with numerical quantification and high throughput operations. To achieve these goals, we developed a modular vector system, pVPX, shown in **Figure 3**.

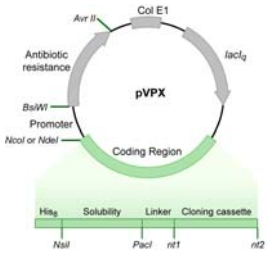

Figure 3. Plasmid map of pVPX, a modularized vector developed at CESG for cell-based protein expression.

This ColE1 replicon provides constitutive expression of the *lac* repressor, variable antibiotic resistance, and options for two different promoters. In addition, the coding region provides a target fusion with His_6 compatible with our downstream purification processes, possibilities for investigating different solubility tags, a variable linker region, and nucleotide sequences (*nt1* and *nt2*) compatible with high throughput cloning and transfer of the cloned and characterized genes.

Table 1. Module variations available in the pVPX vector suite.

Promoter	Antibiotic Resistance	Solubility Tag	Linker region
T5/double *lac* T7/*lac*	kanamycin ampicillin tetracycline	MBP GST thioredoxin NusA	HRV 3C tetraCys

Table 1 summarizes the current set of variations available for the pVPX vector suite. Each of these variations has been engineered to be flanked by unique restriction sites to allow transfers into the appropriate module position. For example, the solubility domain can be exchanged using the *NsiI* and *PacI* restriction sites. Of further interest is the linker region, which encodes the human rhinovirus 3C protease site (HRV) and the tetraCys motif. This linker module facilitates high throughput detection and screening capabilities, as well as proteolytic release of target proteins after purification.

The results presented below were obtained with a T5/amp plasmid giving a His_6/MBP/HRV tetraCys/TEV target fusion as the product. Training and protocol development is currently proceeding with a set of ~100 control proteins whose behaviors were previously characterized by CESG, including results from initial cloning from T87 tissue culture to structure determination.

FLASH Labeling

FLASH labeling was introduced by Tsien and colleagues [1,2]. The tetraCys motif in the linker region between the solubility tag and target protein allows highly specific, covalent labeling of the fusion protein upon reaction with the *bis*-arsenical fluorophore FLASH-EDT_2, as shown in **Figure 4**.

Figure 4. Proteins containing the tetraCys motif can be selectively labeled using *bis*-arsenical fluorophores such as FLASH-EDT_2.

FLASH Labeling and Quantification

Figure 5A shows that unreacted FLASH-EDT_2 has a low fluorescence intensity, while reaction with the tetraCys motif gives rises to a strong fluorescence signal. This large change in fluorescence intensity permits quantification of FLASH-labeled proteins by comparison to a standard curve such as that shown in **Figure 5B**, generated using known concentrations of the control protein His_6-tetraCys-MBP. Based on this analysis, the concentration sensitivity for tetraCys proteins expressed in bacterial cell lysates is better than 5 μM. This routine performance is comparable to the best-case results from Coomassie staining by SDS-PAGE, but can be obtained more rapidly and as a spectral measurement.

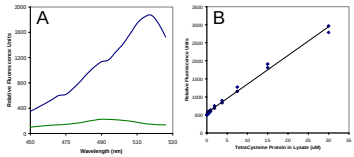

Figure 5. ***A***, increase in fluorescence intensity and shift in the emission maximum for bound FLASH (*blue* line) as compared to free FLASH-EDT_2 (*green* line). ***B***, standard curve generated from reaction of known concentrations of tetraCys fusion protein in a bacterial total cell lysate. The fluorescence spectra and the intensity measurements used to generate the standard curve were obtained in 384-well plates using a Tecan Ultra 384 spectrofluorimeter.

FLASH Labeling and Detection

FLASH labeling provides a diagnostic signal for status of an expressed fusion protein in the bacterial cell lysate. This is demonstrated in **Figure 6**, which shows cell lysates obtained from bacteria expressing a variety of control proteins (C1, C2, and C3) and unknown proteins from *Arabidopsis* (A1-A12). Visual inspection of the Coomassie-stained gel (*top*) reveals a wide range in the level of expression, typical of the results observed in structural genomics work. For the proteins expressed at lower levels, the unambiguous assignment of a lysate band to the fusion protein of interest can be difficult. In contrast, examination of the FLASH-labeled gel *(bottom*) provides a more direct assessment of the presence of the tetraCys motif in the expressed fusion protein.

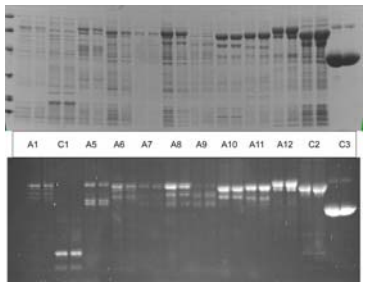

Figure 6. An SDS-PAGE separation of bacterial lysates treated with FLASH-EDT_2 and then visualized using either Coomassie staining (*top*) or fluorescence imaging (*bottom*). Replicates of each sample were analysed. Lanes C1-C3 contain control proteins, lanes A1-A12 contain unknown proteins from *Arabidopsis*. Fluorescence imaging was performed using UV excitation and a FOTODYNE CCD imaging system.

Other Detection Possibilities

Some of the proteins shown in **Figure 6** were expressed at low levels, and FLASH-labeling revealed that truncated versions of the fusion protein represented a significant fraction of the total recombinant protein. The FLASH detection method provides a useful method to unambiguously detect these failures in protein expression, which may arise from a number of biological mechanisms. In this case, the FLASH-labeling provides an assay method to investigate and remediate this problem of faulty expression. FLASH-labeling can be used to investigate the partition of protein between soluble and insoluble fractions upon suitable fractionation of the bacterial cells. FLASH-labeling can also be used to quantitate the amount of protein in column fractions. Each of these applications can be completed on less than 15 μL of sample in high throughput formats using spectrofluorimeters such as the Tecan Ultra 384. Other applications are likely to arise as proteins containing this fusion are evaluated in our protein production pipeline.

Real-time Detection of Proteolysis

Incomplete proteolysis of the target from the fusion protein has been observed with a substantial number of the targets examined at CESG. This may arise if the target has an inaccessible *N*-terminus due to structural considerations or may represent inaccessibility due to misfolding or aggregation of the target protein.

The pVP-X4 vector was designed to allow measurement of both the rate and percentage completion of the proteolysis of fusion proteins [3]. **Figure 7** summarizes how this construct can be used to provide real-time, quantitative analysis of both the rate and the extent of fusion protein proteolysis by the use of fluorescence anisotropy. Incubation of the structural genomics target with the appropriate protease(s) gives rise to a time-dependent decrease in fluorescence anisotropy arising from release of the small, labeled tetraCys peptide from the larger fusion protein. In this way, fluorescent fusion proteins obtained from pVP-X4 yield numerical quantification for total expression, solubility, and protease cleavage from high throughput compatible measurements.

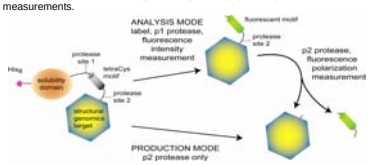

Figure 7. Use of fluorescence polarization to give real time characterization of the rate and extent of the proteolysis of fusion proteins.

In practice, we have found that the rate and extent of proteolysis observed with structural genomics targets is dependent on the nature of the target. **Figure 8** shows fractional cleavage calculated from polarization data for an *Arabidopsis* target that gave ~50% completion of the proteolysis reaction in two hours (*black* circles). In contrast, polarization data for another *Arabidopsis* target (*blue* diamonds) revealed complete proteolysis in the same time period. The assessments made from the fluorescence measurements have been corroborated using SDS-PAGE analysis [3]. In addition to providing useful real-time process information for pipeline production efforts, these measurements also provide opportunities to investigate solution conditions, vector constructs, and protease identity as contributors to improved performance. In other contexts, the availability of real-time assays in a high throughput format can also facilitate the search for inhibitors of protease assays.

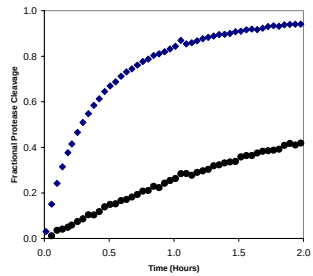

Figure 8. Extent of proteolysis can be determined in real time by monitoring the change in fluorescence anisotropy of FLASH labeled proteins. Cleavage releases a small peptide containing FLASH labeled tetraCys which tumbles at a faster rate than the tetraCys motif attached to the fusion and/or target protein. Shown in *blue* diamonds is the extent of reaction determined for At3g16990.1 proteolysis from MBP. *Black* circles indicate the extent of reaction for removal of At3g03410.1. Both proteins were present at approximately 5 μM with 0.5 μM TEV protease. After two hours, the proteolysis for At3g16990.1 was largely complete while At3g03410.1 was less than 50% complete, indicating that this protein should be incubated for a longer time to effect tag removal.

References

1. B. A. Griffin, S. R. Adams, J. Jones, and R. Y. Tsien, Fluorescent labeling of recombinant proteins in living cells with flash, Methods Enzymol 327 (2000) 565-578.
2. S. R. Adams, R. E. Campbell, L. A. Gross, B. R. Martin, G. K. Walkup, Y. Yao, J. Llopis, and R. Y. Tsien, New biarsenical ligands and tetracysteine motifs for protein labeling in vitro and in vivo: Synthesis and biological applications, J Am Chem Soc 124 (2002) 6063-6076.
3. P. G. Blommel, and B. G. Fox, Fluorescence anisotropy assay for proteolysis of specifically labeled fusion proteins, Anal Biochem 336 (2005) 75-86.

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