Available online at www.sciencedirect.com
ScienceDirect
Protein Expression and Purification 51 (2007) 1–10
Protein Expression & Purification
www.elsevier.com/locate/yprep

Review

Strategies to maximize heterologous protein expression in *Escherichia coli* with minimal cost

Wolfgang Peti [a], Rebecca Page [b,*]

[a] *Brown University, Department of Molecular Pharmacology, Physiology, and Biotechnology, Box G-E3, Providence, RI 02912, USA*
[b] *Brown University, Department of Molecular Biology, Cell Biology and Biochemistry, Box G-E4, Providence, RI 02912, USA*

Received 26 April 2006, and in revised form 20 June 2006
Available online 4 July 2006

Abstract

Automation and miniaturization are key issues of high-throughput research projects in the post-genomic era. The implementation of robotics and parallelization has enabled researchers to process large numbers of protein targets for structural studies in a short time with reasonable cost efficiency. However, the cost of implementing the robotics and parallelization often prohibit their use in the traditional academic laboratory. Fortunately, multiple groups have made significant efforts to minimize the cost of heterologous protein expression for the production of protein samples in quantities suitable for high resolution structural studies. In this review, we describe recent efforts to continue to minimize the cost for the parallel processing of multiple protein targets and focus on those materials and strategies that are highly suitable for the traditional academic laboratory.

Keywords: Structural biology; Heterologous protein expression; *Escherichia coli*; Cloning

Significant amounts of protein, usually between 5 and 50 mg, depending on the protein size and experimental technique used, are required for every structural biology project [1,2], independent of the structure elucidation technique used, including X-ray crystallography, NMR spectroscopy and cryo-electron microscopy. In general, *Escherichia coli* is the preferred host for recombinant protein expression for structural studies [3,4] because it is rather easy to genetically manipulate, it is relatively inexpensive to culture, isotope labeling protocols for NMR spectroscopy and selenomethionine incorporation for X-ray crystallography are established, and expression is fast, typically producing protein in a single day. The importance of *E. coli* for heterologous protein production is perhaps best highlighted by the wide variety of commercial products available for the *E. coli* expression system. However, there are disadvantages to using *E. coli* as an expression host. Namely, *E. coli* is not capable of producing eukaryotic post-translational modifications, such as glycosylation, which can be critical for the production of folded, active protein. Equally important, some proteins, especially larger proteins and membrane proteins, simply fail to express in *E. coli*, or express, but do so insolubly as inclusion bodies.

To overcome some of these limitations, a large number of commercial vectors that facilitate soluble expression and single step purification via the use of different fusion tags have been developed. In addition, multiple *E. coli* strains that facilitate the expression of membrane proteins [5], proteins with rare codons [6], proteins with disulfide bonds [7], proteins that are otherwise toxic to the cell, among others, are readily available. This variety of expression vectors and cell lines now significantly enhances the likelihood of designing an *E. coli* protein expression protocol suitable for the production of the substantial amounts of protein required for structural studies [8]. Finally, recently introduced microexpression incubator shakers that require as little as 500 μl of expression medium enable researchers to screen for optimal expression conditions rapidly with reduced costs [9–11]. Once

* Corresponding author. Fax: +1 401 863 9653.
E-mail address: rebecca_page@brown.edu (R. Page).

1046-5928/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.pep.2006.06.024

the optimal expression construct is identified, standard incubator shakers or fermentors are used to produce large amounts of recombinant overexpressed protein.

During the last few years, both the pharmaceutical industry and the structural genomics community have made significant efforts to develop automated technologies to facilitate the identification and production of proteins suitable for functional and structural studies [12]. This has led to new robotic instrumentation for all steps of the structure determination process, including rapid cloning systems [13], parallel expression and purification technologies [14,15], nanocrystallization [16] and crystal growth imaging, and robotic crystal diffraction screening [17,18]. Within the NIH Protein Structure Initiative, implementation of these technologies has resulted in the determination of more than 1000 new protein structures deposited in the Protein DataBank (PDB; www.nigms.nih.gov/News/Results/021005.htm). However, academic laboratories that usually focus on a few biologically relevant proteins typically do not require the high-throughput demanded by structural genomics centers and pharmaceutical industries. Moreover, most of these technologies, such as robotics, and novel strategies, such as commercially available rapid cloning systems, are cost-prohibitive for most academic laboratories. Instead, cost-effective alternatives are needed. Fortunately, during the last 5 years multiple groups have developed and begun distributing materials that allow for the efficient, parallel screening of multiple constructs with minimal cost. These techniques and materials use standard molecular biology and protein purification instrumentation that can be found in most biology and certainly structural biology laboratories. These new alternatives include new expression vectors [19], new or revitalized fusion tags to facilitate soluble protein expression [20,21], new microexpression/solubility screening protocols [9,22] and new macroexpression methods and instrumentation [23]. In this review, we describe our recent experiences transitioning the parallel approach for protein expression and purification screening commonly used in structural genomics centers to our own academic laboratories and report on those materials and technologies we have found most useful for our own projects.

New approaches to construct identification

One of the most challenging steps in any structural biology project is predicting which protein or protein fragment will express solubly and purify for subsequent NMR spectroscopic or crystallographic studies. Often, small differences in the amino acid sequence itself, or in the length of the construct, can transform a protein that fails to express into one that expresses, purifies and crystallizes readily [24,25]. In spite of multiple efforts to analyze the large amounts of data generated by structural genomics consortia regarding protein expression, purification and crystallization [26,27], we are still unable to predict, based on sequence alone, which proteins will express and purify.

As with most laboratories, combined functional and structural information is used to guide initial attempts to identify the optimal boundaries (starting and ending residues) of a protein target/protein domain. We have found that the 'Fold and Function Assignment System' (FFAS; http://ffas.ljcrf.edu/ffas-cgi/cgi/ffas.pl)[1] is a particularly useful program for identifying weak structural homologs to a sequence of interest. FFAS uses profile-profile sequence alignments and fold recognition to detect remote homologies not identified using other sequence comparison methods [28]. We use FFAS to identify the closest structural homologs to various protein domains of interest to facilitate the identification of appropriate residue boundaries for primer design. When high resolution structures of homologous protein/protein domains are not available, we use secondary structure prediction programs such as PsiPred [29] to identify which regions of the domain are most likely to form stable secondary structural elements. We also use the programs, PONDR [30] and DisEMBL (HotLoops) [31] to predict regions which are disordered. Both PONDR and DisEMBL are computational methods based on neural artificial networks trained for different definitions of disorder, including protein sequences not visible in electron density maps (PONDR) and loops with high B-factors, indicative of a high intrinsic mobility in these regions (DisEMBL). The results from these analyses are then used to identify multiple residues as optimal 'start' and 'end' residues for the constructs. Typically, we select between 2 and 4 'start' residues and 2 and 4 'end' residues and use them combinatorially to subclone anywhere between 4 and 16 different constructs of a single domain for expression, purification and structural analysis.

An additional approach that has proven to be successful for the production of proteins for high resolution studies is ortholog screening [25]. Orthologs are proteins with the same function from different species which, due to evolution, have small differences in their amino acid sequences. While these differences do not affect protein function, they can have significant effects on the ability of a protein to express in *E. coli*, purify and crystallize. Ortholog screening for high resolution studies have been successfully used by multiple groups to obtain protein samples suitable for high resolution studies. For example, within the mouse homolog group of the Joint Center for Structural Genomics (JCSG), 14 proteins were selected for ortholog screening that had either previously failed to crystallize or failed to express in selenomethionine media. Using orthologous proteins, the JCSG was able to solve the high resolution structures of 5 of these 14 protein families. Similarly, the Ontario Center for Structural Genomics found that including even a single ortholog of a target protein increases the number of samples for structural studies by a factor of two [25].

[1] *Abbreviations used:* MCS, multiple cloning sites; TEV, Tobacco Etch Viral; MBP, maltose binding protein; GST, glutathione *S*-transferase; IMAC, immobilized metal affinity chromatography; rpm, revolutions per minute.