Chapter 1

Recombinant Protein Expression in *E. coli*: A Historical Perspective

Opher Gileadi

Abstract

This introductory chapter provides a brief historical survey of the key elements incorporated into commonly used *E. coli*-based expression systems. The highest impact in expression technology is associated with innovations that were based on extensively studied biological systems, and where the tools were widely distributed in the academic community.

Key words *E. coli*, Promoter, Recombinant protein, Protein engineering, Expression vectors

1 Introduction

Early studies on purified proteins depended on proteins found in relatively high abundance, or with distinct solubility and stability profiles, such as hemoglobin, albumin, and casein. Even with the expansion of interest into a wider universe of enzymes, hormones, and structural proteins, researchers have sought to purify proteins from sources (organisms, tissues, and organelles) containing the highest abundance of the desired protein. It was recognized, even before the era of genetic engineering, that microorganisms and cultured cells could be ideal sources for protein production. A remarkable example, just before the development of recombinant DNA technologies, was the overproduction of the lactose repressor (product of the lacI gene). This protein is normally produced in *E. coli* at ~10 copies/cell. Muller-Hill and colleagues [1] used clever selection techniques to isolate promoter mutations that led to a tenfold increase in protein expression; this allele ($lacI^q$) was then transferred to a lysis-deficient bacteriophage, allowing achieving very high copy numbers of the phage (and the $lacI^q$ gene), leading to the target protein being ~0.5 % of total cellular protein [1]; all this—without restriction enzymes and in vitro DNA recombination! The emergence of precision recombinant DNA techniques

Nicola A. Burgess-Brown (ed.), *Heterologous Gene Expression in E. coli: Methods and Protocols*, Methods in Molecular Biology, vol. 1586, DOI 10.1007/978-1-4939-6887-9_1,

led to the production of the first biotechnology-derived drugs, insulin, growth hormone, and interferons, subsequently expanding to 23 FDA-approved biologic drugs produced in *E. coli* [2]. Concurrently, thousands of other proteins were produced in bacteria for research purposes. In this chapter, I will briefly review the major innovations that created the toolkit for recombinant protein expression in *E. coli*.

2 Expression from *E. coli* RNAP Promoters

We have already seen the first principles driving high-efficiency recombinant gene expression: strong promoters, and high gene copy numbers. A third principle that became important early on is inducible gene expression; typically, an expression process will involve growth of cells in the absence of expression, then induction of gene expression through transcriptional regulatory elements or by infection or activation of viruses. Expression vectors were developed based on a small number of well-studied gene promoter systems, which remain popular to this day (reviewed in ref. 3). The Lac promoter/operator and its derivatives (UV5, tac) are inducible by galactose or Isopropyl β-D-1-thiogalactopyranoside (IPTG), and repressed by glucose. The phage lambda P_L promoter is one of the strongest promoters known for *E. coli* RNA polymerase (RNAP). When combined with a temperature-sensitive repressor (cI847), the P_L promoter can be induced by a temperature shift, avoiding the use of chemical inducers [4]. The araBCD promoter, tightly regulated by the araC repressor/activator, avoids leaky expression in the absence of the inducer arabinose [5]. Interestingly, synthetic *E. coli* RNAP promoters based on a consensus derived from multiple sequence alignment perform rather poorly [6, 7]; rather, it is a combination of the canonical −35 and −10 elements with less defined downstream sequences, as well as an optimal environment for protein synthesis initiation and elongation that drives the highest levels of expression.

3 Maximizing Expression Levels

For most applications, *E. coli* RNAP promoters have been superseded by expression systems using bacteriophage promoters and RNA polymerases. The bacteriophage T7 polymerase is highly selective for cognate phage promoters, and achieves very high levels of expression [8]. The commonly used T7 expression systems are regulated by a double-lock: lac operators (repressor-binding sites) are placed at the promoter driving the target gene as well as the promoter driving the expression of the T7 RNA polymerase [9]. Expression is repressed in the absence of inducer, and is rapidly

turned on when IPTG is added. There is some expression in the absence of inducer, which can be further reduced by including glucose in the growth medium (catabolite repression) [10] and by expressing T7 lysozyme, an inhibitor of T7 RNA polymerase, from plasmids pLysS or pLysL [9]. With the successful implementation of these principles, other issues become rate-limiting. High-level expression of foreign genes may be hampered by codon usage that is nonoptimal for the host cell. This makes a real difference [11], and has been addressed using either synthetic, codon-optimized genes, or by co-expressing a set of tRNA molecules that recognize some of the codons that are rare in *E. coli* (available as commercial strains, such as Rosetta™ and CodonPlus). Sequence optimization may also affect other impediments to gene expression, such as mRNA secondary structure or mRNA degradation, as well as secondary advantages such as eliminating or introducing restriction sites.

4 Fusion Tags

The next major development has been the introduction of generic purification tags. The general principle is to genetically fuse the protein of interest to another protein or peptide, for which affinity purification reagents are available. The tags introduced in the late 1980s are still very widely used. The earliest were epitope tags [12]: short peptides that are recognized by monoclonal antibodies, allowing affinity purification and elution with free peptides (e.g., FLAG [13], HA [14], and myc [15] tags). These were followed by the hexahistidine tag [16] which allows purification by immobilized-metal affinity chromatography (IMAC), and the full-length protein glutathione S-transferase (GST) [17] which binds to glutathione-sepharose. Short peptide tags are sometimes concatenated to provide better avidity of binding to the affinity columns, allowing more stringent washes and better purity, but these are mostly used for expression in eukaryotic cells. Tags can be removed using sequence-specific proteases (enterokinase, the blood-clotting factors X and thrombin, viral proteases such as TEV and the rhinovirus 3C protease, SUMO protease, engineered subtilisin, or inteins). Fusion tags seem to perform at least two functions: first, providing a handle for affinity purification; and second, promoting the solubility of the target protein by changing the overall hydrophobicity and charge and by providing chaperone-like functions. Because the selectivity and the solubilizing effect are context-dependent, there has been a continuing development of new fusion tags to address specific goals in different cell types.

It is frequently observed that the highest expression levels of a recombinant protein do not necessarily correlate with the highest yields of soluble, properly folded protein. In fact, rapid production

of heterologous proteins more often leads to aggregation and precipitation, with no recovery of active protein. This problem has been addressed using three approaches: modulating growth and induction conditions; modifying the host strain; and engineering the target protein. Many eukaryotic proteins expressed in *E. coli* are only soluble when induced at low temperatures, typically 15–25 °C. Other changes in induction conditions, such as the use of carefully calibrated autoinduction media [10] and the use of moderately active promoters, have on occasion led to higher yields. Host strains have been engineered to over-express chaperone proteins [18–20], to encourage disulfide bond formation [21], or to remove autophosphorylated sites from active protein kinases [22]. Finally, proteins can be recovered from denatured precipitates using refolding techniques following solubilization in guanidine or urea; however, refolding methods seem to be mostly effective only for a subset of proteins, predominantly extracellular domains or proteins. The recent application of high-throughput and design of experiment methods to optimize refolding conditions may help to rescue more proteins that cannot be properly folded during expression in bacteria.

5 The Protein Is the Most Important Variable

The most dramatic improvements in recovery of soluble proteins have come from optimizing the sequence of the expressed protein. The degree of flexibility in the engineering of the target protein depends on the purpose of the project. In many cases, a truncated protein that contains a well-folded globular domain will be solubly expressed, while the full-length protein may contain intrinsically disordered and hydrophobic regions that drive aggregation. This is particularly relevant when expressing proteins for crystallization, and it has been noted that constructs truncated to include the structured domains tend to express and crystallize well [23]. In addition to truncations, internal mutations that stabilize the protein can dramatically affect the yields of soluble proteins [24] as well as membrane proteins [25, 26]; identifying these mutants most often requires molecular evolution techniques, as there is rarely any solid basis for rational design, especially if the structure of the protein is unknown. A more natural version relies on natural diversity: very often, systematic cloning and test-expression of multiple orthologues of the target protein can lead to the identification of a related protein that does express well in *E. coli*. Alternatively, synthetic versions of the target proteins based on multiple sequence alignments have been used in some instances to generate better yields.

6 High-Throughput Methods

With the advent of genomic-scale studies, there was a need to streamline and parallelize the cloning process. New methods were developed to enable cloning of PCR-generated DNA fragments into vectors without prior cleavage by restriction enzymes, and cloning of each fragment into multiple vectors. These methods include variants of ligation-independent cloning (LIC) [23, 27–29] and site-specific recombination methods [30]. The choice of method depends on the details of the experimental goals: LIC methods require only minimal (or no) additions to the cloned sequence, while recombinase-based methods (e.g., the Gateway® method) [30] add obligatory sequences within the encoded protein. On the other hand, when there is a need to repeatedly clone the same fragment into multiple vectors, recombinase-based methods allow a sequence-verified DNA insert to be transferred in a virtually non-mutagenic manner. An additional development to enable efficient cloning with low background has been the introduction of toxic genes in cloning vectors that are inactivated by the insertion of the cloned fragments [31, 32].

7 Heteromeric Complexes

It has been realized for a long time that attempts to express individual polypeptides in heterologous cells may fail because the native structure of the protein requires hetero-oligomerization. Techniques for co-expression of several components of a protein complex were applied sporadically, combining more than one protein/transcription unit on a single plasmid, or by combining separate compatible plasmids in a bacterial cell (or a combination of both). Recently developed systems for recombining multiple coding sequences into one plasmid [33] will allow generating protein complexes efficiently and systematically in *E. coli*.

8 One Method Fits All?

A search of GenBank for organism/vector yields >8000 hits; it would be safe to estimate the number of *E. coli* expression vectors is at least 1000. There are probably $>10^4$ publications describing the expression and purification of individual proteins, all differing at least slightly in the experimental details; the information is very difficult to collate. The structural genomics projects in the US, Europe, and Japan have systematically expressed and purified proteins from a variety of organisms, with extensive documentation

and several benchmarking studies to evaluate the success of different approaches. A paper published jointly in 2008 by most of the big players [34] shows that a fairly narrow range of techniques accounts for the vast majority of successfully produced proteins. Some more detailed comparative studies (e.g., [29, 35]) have shown that by far the most common combination is BL21(DE3)-derived host strains supplemented with rare-codon tRNAs; growth in rich medium, with either IPTG-driven or autoinduction at 20–25 °C. The biggest impact on the yield of soluble protein is linked to (1) construct selection (truncation/mutation); (2) fusion tags, and (3) lowering the temperature during induction. Do these statistics mean that more than 35 years of method development is almost redundant, beyond a handful of core methods that cover all our needs? Probably not; the aggregate statistics hide the fact that the parameters of the structural genomics projects allowed for a considerable failure rate; in practice, the core methods (and the variants used) could recover soluble proteins for less than 50 % of eukaryotic target proteins that were attempted. Individual proteins may be rescued by more sophisticated solutions developed over the years, as documented in this volume. However, it is likely that these methods will have a marginal effect on the overall success rates of expressing eukaryotic proteins in *E. coli*, leaving us with a sizeable fraction of proteins that cannot be productively expressed.

9 Future Prospects

What are the future prospects? On one hand, it is sensible to transfer proteins that consistently fail to be produced in *E. coli* to other expression systems, which are becoming more efficient and cost-effective. However, it is likely that bacteria will continue to be a major workhorse for recombinant protein expression. One point that emerges from this historical survey is that most significant developments were based on thorough knowledge of particular biological systems. Indeed, the choice of *E. coli* and Coliphage-derived elements was a consequence of decades of fundamental research on these organisms, starting from the 1940s [36]. A recent splendid example of the use of in-depth fundamental research is the development of CRISPR-Cas9 systems for gene editing [37, 38]. So, true innovation in expanding the universe of proteins that can be produced in bacterial cells is likely to come from unexpected areas, based on in-depth knowledge. I would hazard a guess that big developments will come from synthetic biology. The engineering of *E. coli* host strains has proceeded piecemeal, typically adding or modifying individual proteins or pathways [39, 40]. Yet, a variety of other bacteria are used as host strains, including *Pseudomonas* and *Bacillus subtilis*, which provide specific advantages. With the advent of fully engineered bacterial

cells [41] and the reconstitution of complex metabolic pathways [42, 43], it is plausible that novel "protein factories" will be designed to incorporate features from a variety of expression systems, to provide features that are missing or suboptimal in current *E. coli* hosts. These may include posttranslational modifications, chaperone functions, incorporation into membranes with controllable lipid composition, and secretion to the culture media. Parallel efforts will include extensive protein evolution to derive well-behaved and highly expressed versions of the proteins of interest.

As a final note, it is maybe obvious that the most widely adapted techniques and expression systems are those that were widely available to the academic community (at least), either through open distribution (by organizations such as Addgene [44]) or through reasonably priced vendors. It is imperative that future core technologies are not protected to an extent that makes them practically inaccessible to the majority of researchers. A sensible mix of commercial licensing and academic freedom-to-operate can benefit both the inventors and the society at large.

References

1. Muller-Hill B, Crapo L, Gilbert W (1968) Mutants that make more lac repressor. Proc Natl Acad Sci U S A 59:1259–1264
2. Baeshen MN, Al-Hejin AM, Bora RS et al (2015) Production of biopharmaceuticals in *E. coli*: current scenario and future perspectives. J Microbiol Biotechnol 25:953–962
3. Baneyx F (1999) Recombinant protein expression in *Escherichia coli*. Curr Opin Biotechnol 10:411–421
4. Remaut E, Stanssens P, Fiers W (1983) Inducible high level synthesis of mature human fibroblast interferon in *Escherichia coli*. Nucleic Acids Res 11:4677–4688
5. Guzman LM, Belin D, Carson MJ et al (1995) Tight regulation, modulation, and high-level expression by vectors containing the arabinose PBAD promoter. J Bacteriol 177:4121–4130
6. Brunner M, Bujard H (1987) Promoter recognition and promoter strength in the *Escherichia coli* system. EMBO J 6:3139–3144
7. Deuschle U, Kammerer W, Gentz R et al (1986) Promoters of *Escherichia coli*: a hierarchy of in vivo strength indicates alternate structures. EMBO J 5:2987–2994
8. Rosenberg AH, Lade BN, Chui DS et al (1987) Vectors for selective expression of cloned DNAs by T7 RNA polymerase. Gene 56: 125–135
9. Dubendorff JW, Studier FW (1991) Controlling basal expression in an inducible T7 expression system by blocking the target T7 promoter with lac repressor. J Mol Biol 219:45–59
10. Studier FW (2014) Stable expression clones and auto-induction for protein production in *E. coli*. Methods Mol Biol 1091:17–32
11. Burgess-Brown NA, Sharma S, Sobott F et al (2008) Codon optimization can improve expression of human genes in *Escherichia coli*: a multi-gene study. Protein Expr Purif 59: 94–102
12. Munro S, Pelham HR (1984) Use of peptide tagging to detect proteins expressed from cloned genes: deletion mapping functional domains of *Drosophila* hsp 70. EMBO J 3: 3087–3093
13. Hopp TP, Prickett KS, Price VL et al (1988) A short polypeptide marker sequence useful for recombinant protein identification and purification. Nat Biotechnol 6:1204–1210
14. Field J, Nikawa J, Broek D et al (1988) Purification of a RAS-responsive adenylyl cyclase complex from *Saccharomyces cerevisiae* by use of an epitope addition method. Mol Cell Biol 8:2159–2165
15. Robertson D, Paterson HF, Adamson P et al (1995) Ultrastructural localization of ras-related proteins using epitope-tagged plasmids. J Histochem Cytochem 43:471–480
16. Hochuli E, Dobeli H, Schacher A (1987) New metal chelate adsorbent selective for proteins and peptides containing neighbouring histidine residues. J Chromatogr 411:177–184

17. Smith DB, Johnson KS (1988) Single-step purification of polypeptides expressed in *Escherichia coli* as fusions with glutathione S-transferase. Gene 67:31–40
18. Lee SC, Olins PO (1992) Effect of overproduction of heat shock chaperones GroESL and DnaK on human procollagenase production in *Escherichia coli*. J Biol Chem 267:2849–2852
19. Nishihara K, Kanemori M, Kitagawa M et al (1998) Chaperone coexpression plasmids: differential and synergistic roles of DnaK-DnaJ-GrpE and GroEL-GroES in assisting folding of an allergen of Japanese cedar pollen, Cryj2, in *Escherichia coli*. Appl Environ Microbiol 64: 1694–1699
20. Ferrer M, Chernikova TN, Timmis KN et al (2004) Expression of a temperature-sensitive esterase in a novel chaperone-based *Escherichia coli* strain. Appl Environ Microbiol 70: 4499–4504
21. Bessette PH, Aslund F, Beckwith J et al (1999) Efficient folding of proteins with multiple disulfide bonds in the *Escherichia coli* cytoplasm. Proc Natl Acad Sci U S A 96:13703–13708
22. Shrestha A, Hamilton G, O'Neill E et al (2012) Analysis of conditions affecting autophosphorylation of human kinases during expression in bacteria. Protein Expr Purif 81:136–143
23. Savitsky P, Bray J, Cooper CD et al (2010) High-throughput production of human proteins for crystallization: the SGC experience. J Struct Biol 172:3–13
24. Tsai J, Lee JT, Wang W et al (2008) Discovery of a selective inhibitor of oncogenic B-Raf kinase with potent antimelanoma activity. Proc Natl Acad Sci U S A 105:3041–3046
25. Schlinkmann KM, Hillenbrand M, Rittner A et al (2012) Maximizing detergent stability and functional expression of a GPCR by exhaustive recombination and evolution. J Mol Biol 422:414–428
26. Serrano-Vega MJ, Magnani F, Shibata Y et al (2008) Conformational thermostabilization of the beta1-adrenergic receptor in a detergent-resistant form. Proc Natl Acad Sci U S A 105: 877–882
27. Aslanidis C, de Jong PJ (1990) Ligation-independent cloning of PCR products (LIC-PCR). Nucleic Acids Res 18:6069–6074
28. Klock HE, Lesley SA (2009) The polymerase incomplete primer extension (PIPE) method applied to high-throughput cloning and site-directed mutagenesis. Methods Mol Biol 498:91–103
29. Unger T, Jacobovitch Y, Dantes A et al (2010) Applications of the restriction free (RF) cloning procedure for molecular manipulations and protein expression. J Struct Biol 172:34–44
30. Hartley JL, Temple GF, Brasch MA (2000) DNA cloning using in vitro site-specific recombination. Genome Res 10:1788–1795
31. Bernard P, Gabant P, Bahassi EM et al (1994) Positive-selection vectors using the F plasmid ccdB killer gene. Gene 148:71–74
32. Gay P, Le Coq D, Steinmetz M et al (1985) Positive selection procedure for entrapment of insertion sequence elements in gram-negative bacteria. J Bacteriol 164:918–921
33. Haffke M, Marek M, Pelosse M et al (2015) Characterization and production of protein complexes by co-expression in *Escherichia coli*. Methods Mol Biol 1261:63–89
34. Structural Genomics C, China Structural Genomics C, Northeast Structural Genomics C et al (2008) Protein production and purification. Nat Methods 5:135–146
35. Vincentelli R, Cimino A, Geerlof A et al (2011) High-throughput protein expression screening and purification in *Escherichia coli*. Methods 55:65–72
36. Cairns J, Stent GS, Watson JD (2007) In: Centennial (ed) Phage and the origins of molecular biology. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY
37. Cong L, Ran FA, Cox D et al (2013) Multiplex genome engineering using CRISPR/Cas systems. Science 339:819–823
38. Mali P, Yang L, Esvelt KM et al (2013) RNA-guided human genome engineering via Cas9. Science 339:823–826
39. Chen R (2012) Bacterial expression systems for recombinant protein production: *E. coli* and beyond. Biotechnol Adv 30:1102–1107
40. Makino T, Skretas G, Georgiou G (2011) Strain engineering for improved expression of recombinant proteins in bacteria. Microb Cell Fact 10:32
41. Hutchison CA 3rd, Chuang RY, Noskov VN et al (2016) Design and synthesis of a minimal bacterial genome. Science 351: aad6253
42. Galanie S, Thodey K, Trenchard IJ et al (2015) Complete biosynthesis of opioids in yeast. Science 349:1095–1100
43. Paddon CJ, Westfall PJ, Pitera DJ et al (2013) High-level semi-synthetic production of the potent antimalarial artemisinin. Nature 496:528–532
44. Kamens J (2015) The Addgene repository: an international nonprofit plasmid and data resource. Nucleic Acids Res 43:D1152–D1157