Recombinant protein expression in *Escherichia coli*
François Baneyx

*Escherichia coli* is one of the most widely used hosts for the production of heterologous proteins and its genetics are far better characterized than those of any other microorganism. Recent progress in the fundamental understanding of transcription, translation, and protein folding in *E. coli*, together with serendipitous discoveries and the availability of improved genetic tools are making this bacterium more valuable than ever for the expression of complex eukaryotic proteins.

Addresses
Department of Chemical Engineering, University of Washington, Box 351750, Seattle, WA 98195, USA; e-mail: baneyx@cheme.washington.edu

Current Opinion in Biotechnology 1999, 10:411–421

© 1999 Elsevier Science Ltd.

Introduction
Among the many systems available for heterologous protein production, the Gram-negative bacterium *Escherichia coli* remains one of the most attractive because of its ability to grow rapidly and at high density on inexpensive substrates, its well-characterized genetics and the availability of an increasingly large number of cloning vectors and mutant host strains. Although there is no guarantee that a recombinant gene product will accumulate in *E. coli* at high levels in a full-length and biologically active form, a considerable amount of effort has been directed at improving the performance and versatility of this workhorse microorganism. This review will examine the salient features of *E. coli*-based expression systems with an emphasis on how a number of limitations have recently been addressed. Many additional details and references can be found in Makrides' excellent review [1].

Plasmid copy number and maintenance
To achieve high gene dosage, heterologous cDNAs are typically cloned into plasmids that replicate in a relaxed fashion and are present at 15–60 (e.g. pMB1/ColE1 derivatives) or a few hundred copies per cell (e.g. the pUC series of pMB1 derivatives). When co-overexpression of additional gene products is desired, ColE1 derivatives are usually combined with compatible plasmids bearing a p15A replicon and maintained at about 10–12 copies per cell. Under laboratory conditions, such multicopy plasmids are randomly distributed during cell division and, in the absence of selective pressure, are lost at low frequency (10⁻⁵–10⁻⁶ per generation) primarily as a result of multimerization [2]. Nevertheless, plasmid loss can increase tremendously in the case of very high copy number plasmids, when plasmid-borne genes are toxic to the host or otherwise significantly reduce its growth rate, or when cells are cultivated at high density or in continuous processes.

The simplest way to address this problem is to take advantage of plasmid-encoded antibiotic-resistance markers and supplement the growth medium with antibiotics to kill plasmid-free cells. The drawbacks of this approach are loss of selective pressure as a result of antibiotics degradation, inactivation, or leakage of periplasmic detoxifying enzymes (e.g. β-lactamase) into the growth medium, and the contamination of the product or biomass by antibiotics, which may be unacceptable from a medical or regulatory standpoint.

A number of alternative strategies have therefore been developed to ensure that plasmid-free cells will not overtake a culture (Table 1). In most cases, cloning vectors are engineered to carry gene(s) or repressors that cause cell death upon plasmid loss. Although all these approaches have proved valuable, they may place restrictions on the growth medium composition in the case of complementation, and introduce a metabolic burden on the cell by requiring transcription, and often translation, of additional plasmid-encoded genes. To circumvent these problems, Williams et al [3] created a host strain containing a conditional essential gene under control of the lac operator/promoter region and a companion multicopy plasmid bearing the lac operon. Titration of the LacI repressor protein by plasmid-encoded lac operators resulted in the expression of the chromosomal gene (in this case kanamycin resistance) and the selective growth of plasmid-bearing cells in medium supplemented with the antibiotic. Replacement of the kanamycin resistance cartridge by an essential host gene will improve the value of this system by removing the disadvantages associated with the use of antibiotic resistance.

A radically different solution to the problem of plasmid instability is the direct insertion of heterologous genes within the chromosome of *E. coli*. Although simple delivery vehicles (e.g. bacteriophage λ) are available for this purpose, little emphasis has been placed on this strategy owing to the perceived notion that gene dosage will necessarily be...
Complementation
An essential chromosomal gene is deleted or mutated and an intact copy or a suppressor is supplied in trans on a plasmid. Plasmid loss leads to cell death under non-permissive growth conditions. Examples of chromosomal alterations include deletions of genes necessary for the synthesis of essential amino acids, and thermosensitive and nonsense mutations in essential chromosomal genes.

Promoters
For many years the E. coli lactose utilization (lac) operon has served as one of the paradigms of prokaryotic regulation. It is therefore not surprising that many of the promoters used to drive the transcription of heterologous genes have been constructed from lac-derived regulatory elements. Although the lac promoter and its close relative, lacUV5 (which is theoretically not subject to cAMP-dependent regulation, but see [17]), are rather weak and rarely used for the high-level production of recombinant polypeptides, they are extremely valuable tools to achieve graded expression of helper or toxic proteins provided that lacY mutant hosts are used and that induction is performed with the non-hydrolyzable lactose analog isopropyl-β-D-1-thiogalactopyranoside (IPTG) (see [9•,8•] for a discussion). The synthetic tac and tr promoters, which consist of the −35 region of the trp promoter and the −10 region of the lac promoter, only differ by 1 bp in the length of the spacer domain separating the two hexamers. Both promoters are quite strong and routinely allow the accumulation of polypeptides to about 15–30% of the total cell protein. Although it is often argued that the cost of IPTG limits the usefulness of these promoters, this is rarely a problem for high-added-value products. Furthermore, as little as 50–100 µM IPTG is usually sufficient to achieve full induction. The more serious issue of IPTG toxicity can be circumvented by utilizing lactose as an inducer or by making use of thermosensitive variants of the LacI repressor protein that allow thermal induction of recombinant protein synthesis.

The weakness of lac-derived promoters may be a concern for the production of membrane proteins or other gene products that are toxic to the cell. For medium copy number plasmids (e.g. pBR322), repression can be efficiently achieved by using host strains carrying the lacI UV5 null allele. This single nucleotide mutation in the −35 hexamer of the chromosomal lacI promoter leads to an increase in the number of LacI repressor molecules from 10–20 to over 100 per cell. For higher copy number plasmids (e.g. pUC derivatives or pMB1 derivatives containing a ram/nop mutation), the lacI Q allele efficiently represses the expression plasmid or provided in trans on a compatible plasmid. It was recently shown, however, that a 15 bp deletion in the lacI promoter that fortuitously replaces the native −35 hexamer by the consensus sequence for β-dependent promoters increases the strength of the lacI promoter 170-fold [10•]. Strains bearing the resulting lacI Q promoter efficiently express lac-regulated genes on high copy number plasmids and full activation of plasmid-borne tac promoters can be achieved with as little as 4–10 µM IPTG [10•].

In recent years, the pET vectors (commercialized by Novagen, Madison, WI) have gained increasing popularity. In this system, target genes are positioned downstream of the bacteriophage T7 late promoter on medium copy number plasmids. The highly processive T7 RNA polymerase...
is supplied in *E. coli*. Typically, production hosts contain a prophage (ophage λDE3) encoding the enzyme under control of the IPTG-inducible *lac*UV5 promoter. While this system leads to the synthesis of large amounts of mRNA, and, in most cases, the concomitant accumulation of the desired protein at very high concentrations (40-50% of the total cell protein), it is not without drawbacks. For example, high levels of mRNA can cause ribosome destruction and cell death, and leaky expression of T7 RNA polymerase may result in plasmid or expression instability. Furthermore, even ‘empty’ pET plasmids are toxic to *E. coli* in the presence of IPTG [11]. Some of the strategies that have been developed to address these issues are co-overexpression of phage T7 lysozyme (which degrades T7 RNA polymerase) from the compatible plasmids and plasmids (Novagen) and the insertion of a lac operator sequence downstream of plasmid-encoded T7 promoters, in order to reduce leaky transcription. In addition, empirical selection has yielded strains that are superior to the traditional BL21(DE3) host by overcoming toxic effects associated with the overexpression of membrane and globular proteins under T7 transcriptional control [11]. Finally, it has been reported that the *lac*UV5 promoter becomes activated in stationary phase cultures in a process requiring cAMP; cAMP-deficient (*cya*) mutants of BL21(DE3) should be used for clone selection and fermentation to avoid counter-selection of plasmids carrying toxic genes under T7-control [7].

An additional limitation of the T7 and other strong promot- er systems is that the target protein is often unable to reach a native conformation, partially or completely segregating within inclusion bodies. Although this problem may be addressed by co-overexpressing folding modulators or through fusion protein technology (see below), an alter- native strategy is to use promoters that are activated by temperature downshift, as proper protein folding is often favored under low temperature cultivation conditions (see [12] and references therein). The best characterized cold shock promoter is that of the major *E. coli* cold-shock protein *CspA* [13•]. Although the *cspA* core promoter is only weakly induced under cold conditions, the upstream region of the gene (−59 NNAAA[A/T][A/T][A/T]TTTTNNAANNN) is strongly regulated by a cold-modified translational machinery, containing fewer polyamines and a larger number of monomers, 308 and 508 ribosomes. The *cspA* promoter is rather well repressed at and above 37°C, compares favorably to the tac promoter for the expression of an aggregation-prone fusion protein at reduced temperatures and remains functional at 10°C [14]. The major disadvantage of the *cspA* system is that it becomes strongly induced 1–2 hours after temperature downshift, a time period that is too short to allow high-level accumulation of recombinant proteins. However, the use of a host strain carrying a null mutation in *rpoA*, a gene encoding a 15 kDa protein associating with free 308 ribosomal subunits, allowed continuous *cspA*-driven production of β-galactosidase in high density cell fermentations for several hours following trans- fer to low temperatures [15•]. The recent demonstration that *cspA*-driven transcription is beneficial for the expres- sion of toxic and proteolytically sensitive gene products, together with the availability of cloning vectors designed for rapidly positioning cDNAxs under *cspA* transcriptional control (M Mujacic, K Cooper, F Baneyx; unpublished data) should stimulate interest in this system. Interestingly, the strong bacteriophage λ*P* promoter, which is typically used to drive the synthesis of recombinant proteins by transferring strains containing a thermostable variant of the λ*P* repressor protein (λ*P*) from 30 to 42°C, is also cold-inducible [16]. In this case, the main drawback is a high basal level of expression as low-temperature induction must be performed in strains lacking λ*P*.

Among the various nutritionally inducible promoters (e.g., pbcl and try, which are induced by phosphate and tryptophan limitation, respectively), the arabinose promot- er (araBAD or PBAD) has recently become commercialized by Invitrogen Corp (Carlsbad, CA). This system uses the inexpensive sugar l-arabinose as an inducer and is somewhat weaker than the tac promoter. Although it is commonly believed that *araBAD* can be used to achieve graded levels of protein expression by varying the arabinose concentration, there is extensive heterogeneity in cell pop- ulations treated with saturating concentrations of the inducer, with some bacteria fully induced and others not at all [9]. Thus, *araBAD* will not be useful for precisely con- trolling the levels of protein accumulation until a host strain that efficiently uptakes arabinose by constitutively synthesizing the arabinose transporter(s), or a gratuitous inducer that does not employ them is identified [8•,9]. Additional promoters regulated by a variety of signals (pH, dissolved oxygen concentration, osmolarity, etc.) are available and have been reviewed in detail elsewhere [17].

**Upstream elements**

The DNA regions that flank core promoters play an important role in determining transcription efficiency. Upstream (UP) elements located 5′ of the −35 hexamer in certain bacterial promoters are A+T rich sequences that increase transcription by interacting with the α subunit of RNA polymerase [18•]. Because few UP elements have been isolated, Gourse and co-workers [19••] used *in vitro* selection to identify upstream sequences conferring increased activity to the *SacB* P1 core promoter. The best UP sequence was portable and increased *in vivo* transcription from the *SacB* P1 and *lac* core promoters 32- and 108-fold, respectively.

The degree of homology with the deduced consensus sequence (−59 NNAAA[A/T][A/T][A/T]TTTTNNAANNN−38 where N is any nucleotide) was also shown to correlate with the strength of natural UP elements fused upstream of the *lac* core promoter [20•]. These results suggest that the
positioning of highly active UP sequences upstream of well
repressed promoters may increase their strength to a level
only achieved thus far with phage promoters, but without
the drawbacks associated with phage polymerase expres-
sion (e.g. leakiness, toxicity and counter-selection).

**mRNA stability**

*E. coli* mRNAs are rather unstable, with half-lives ranging
between 30 and 20 min. The major enzymes involved in
mRNA degradation are two 3′→5′ exonucleases (RNase II
and polynucleotide phosphorylase [PNPase]) and the
endonuclease RNase E. [23] The catalytic activity of
RNase E is located at the protein amino terminus, whereas
the carboxy-terminus serves as a scaffold for the assembly
of a highly efficient RNA 'degradasome' involving PNPase,
the DEAD-box RNA helicase RhlB and the glycolytic
ezyme enolase. There is considerable controversy over
whether RNase E-dependent mRNA decay proceeds in
the 5′→3′ or in the opposite direction. In either case, stable
secondary structures present in the 5′ UTR of certain tran-
scripts as well as in 3′ rho-independent terminators can both
increase mRNA stability; however, their efficiency is
modulated by fine features. For example, addition of
poly(A) tails to the 3′ end of mRNAs by the seemingly
redundant poly(A) polymerases PAP I (the *rho* gene prod-
uct) and PAP II [23] provides a single stranded 'toehold' for
RNase II and PNPase that facilitates transcript degrada-
tion. In general, polyadenylation is not a problem for recomb-
inant mRNAs in wild-type *E. coli* strains.

The stabilizing effect conferred by untruncated 5′ hairpins
was first demonstrated in the case of the long-lived *ompA*
mRNA. Fusion of the *ompA* 5′ UTR to a variety of het-
erogeneous mRNAs significantly increased transcript half-life in
prokaryotic *E. coli* systems [24]. This protective effect is abrogated, however, when the hairpin is preceded by 5′ unpaired nucleotides [25].

Because of their importance in stabilizing sub-
strates with 5′ monophosphate ends than 5′ triphosphate
ends [26••], the stabilizing function of 5′ hairpins may be related to their ability to sequencer the end of the transcript.

The 5′ UTR of the *ompA* mRNA appears particularly well
suited for this task as among 10 synthetic hairpins, only
one was slightly more effective than the *ompA* UTR in sta-
bilizing lacZ transcripts [25].

**Translational issues**

Initiation of translation of *E. coli* mRNAs requires a Shine-
Dalgamo (SD) sequence complementary to the 3′ end of
the 16S rRNA and of consensus 5′-AGGAGG-3′, fol-
lowed by an initiation codon, which is most commonly
AUG. About 8% of start sites use GUG, whereas UUG and
AUU are rare initiators that are only present in autogene-
ously regulated genes (e.g. those encoding ribosomal
protein N20 and initiation factor 3). Although the optimal
spacing between these two features is 8 nt, translation ini-
tiation is only severely affected if this distance is reduced
below 4 nt or increased above 14 nt [27]. Because of the
close coupling between transcription and translation in
prokaryotes, engineering of the translation initiation region
is a powerful tool for modulating gene expression in a pro-
karyote-independent fashion [27]. This also means that
stable mRNA secondary structures encompassing the SD
sequence and/or the initiation codon can dramatically
reduce gene expression by interfering with ribosome bind-
ing. This problem can be circumvented by increasing the
homology of SD regions to the consensus, and by raising
the number of A residues in the initiation region through
directed mutagenesis. An additional mRNA feature
affecting translation initiation is the downstream box (DB),
which is located after the initiation codon and comple-
mentary to bases 1469-1483 of the 16S rRNA. DBs have a
5′-AUGAUUCGCAAAGUG-3′ consensus sequence and
recent evidence suggests that they play a major role as
translational enhancers [28••]. Although introduction of a
consensus DB at the 5′ end of genes encoding recombi-
nant proteins would change their amino acid sequence,
increasing the homology of this region to that of a DB by
using synonymous codons may improve translation initia-
tion of certain transcripts.

Differences in codon usage between prokaryotes and
eukaryotes can have a significant impact on heterologous
protein production. The arginine codons AGA and AGG
are rarely found in *E. coli* genes, whereas they are common in
*Saccharomyces cerevisiae* and eukaryotes. The pro-
tant, but much less obvious effect of AGA codons, is
primary structure changes due to the misincorporation of
lysine for arginine, particularly when cells are grown in
minimal medium [30]. Fortunately, these problems can
usually be addressed by using site-directed mutagenesis to
replace rare arginine codons by the *E. coli*-preferred GUG
codon or by overexpressing the *argU* gene which
encodes an AGA-dependent endonuclease RNase E [21].

**Folding in the cytoplasm**

Overproduction of heterologous proteins in the cytoplasm
of *E. coli* is often accompanied by their misfolding and se-
gregation into insoluble aggregates known as inclusion
bodies. Although inclusion body formation can greatly
simplify protein purification, there is no guarantee that the
in *civo* refolding will yield large amounts of biologically
active product (unsuccessful refolding attempts are seldom
reported in the literature). A traditional approach to reduce
protein aggregation is through fermentation engineering,
most commonly by reducing the cultivation temperature
(see [12] and references therein). The more recent real-
ization that in *civo* protein folding is assisted by molecular
chaperones, which promote the proper isomerization and
cellular targeting of other polypeptides by transiently
interacting with folding intermediates, and by foldases,
which accelerate rate-limiting steps along the folding
The best characterized molecular chaperones in the cytoplasm of E. coli are the ATP-dependent DnaK-DnaJ-GrpE and GroEL-GroES systems [32•,33•]. DnaK binds to hydrophobic regions exposed to the solvent by nascent or stress-unfolded polypeptides, thereby preventing off-pathway reactions leading to aggregation. The promiscuity of DnaK binding is well explained by the fact that it recognizes heptameric stretches of amino acids consisting of a 4–5 residues-long hydrophobic core flanked by basic residues. This motif occurs every 36 residues on the average protein [34]. DnaJ, which independently binds folding intermediates, activates DnaK for tight substrate binding and might direct it to high-affinity sites. The nucleotide exchange factor GrpE mediates complex resolution; released proteins may either fold into a proper conformation, be recaptured by DnaK-DnaJ for additional cycles of interaction or be reversibly transferred to the ‘downstream’ GroEL-GroES chaperonins. GroEL is an ~800 kDa hollow toroid consisting of two stacked homohexameric rings. It binds both substrate proteins and GroES (a 70 kDa dome-shaped homohexamer) via a ring of hydrophobic residues located in its apical domain. Although no clear consensus sequence has been identified, GroEL, like DnaK, appears to favor hydrophobic and basic residues in its substrates [35]. Upon GroES binding, partially structured folding intermediates are released into the inner cavity of GroEL, where they can fold in a capped and hydrophilic environment. There is extensive evidence that co-overproduction of DnaK-DnaJ and GroEL-GroES can greatly increase the soluble yields of aggregation-prone proteins (see [31] and references therein) and a number of plasmids compatible with pMB1-derived cloning vectors are available for this purpose [30•,37,38]. The process does not involve dissolution of preformed recombinant inclusion bodies but is related to improved folding of newly synthesized protein chains [38]. It is important to point out, however, that the beneficial effect associated with an increase in the intracellular concentration of DnaK-DnaJ and GroEL-GroES is highly dependent on the nature of the overproduced protein, and that success is by no means guaranteed (and highly unlikely if the protein is inherently incapable of folding).

Based on in vitro studies and homology considerations, a number of additional cytoplasmic proteins have been proposed to function as molecular chaperones. They include ClpB, HspG and IbpA/B, which, like DnaK-DnaJ-GrpE and GroEL-GroES, are heat-shock proteins (Hsps) belonging to the σ32 stress regulon. Although inactivation of these Hsps has a modest effect on the ability of E. coli to handle thermal stress [39], they appear to have a supporting role in cellular protein folding by acting as minor chaperones that bind folding intermediates or misfolded proteins and transfer them to the DnaK-DnaJ-GrpE team for subsequent reactivation (Figure 1). Although overproduction of IbpA/B, HspG or ClpB did not suppress the misfolding of an aggregation-prone fusion protein in the E. coli cytoplasm (JG Thomas, F Baneyx, unpublished data), increased intracellular levels of these Hsps might improve the solubility of other substrates, particularly if coordinated with DnaK-DnaJ overexpression.

The trans conformation of X–Pro bonds is energetically favored in nascent protein chains; however, ~5% of all prolyl peptide bonds are found in a cis conformation in intracellular proteins. The trans to cis isomerization of X–Pro bonds is rate limiting in the folding of many polypeptides and is catalyzed in vitro by peptidyl prolyl cis/trans isomerases (PPIases). Three cytoplasmic PPIases, SlyD, SlyA and...
trigger factor (TF), have been identified to date [40,41]. The most potent is TF, a 48 kDa protein associated with 50S ribosomal subunits that has been postulated to cooperate with chaperones to guarantee proper folding of newly synthesized proteins. Whether TF overproduction will improve the folding of recombinant proteins synthesized in the E. coli cytoplasm remains to be determined. It should be noted, however, that this may not be without physiological consequences as TF and NidD overproduction lead to cell filamentation. Interestingly, co-overproduction of a leader-less version of PpiA (thus confining to the cytoplasm a PPIase that normally resides in the periplasm) has been shown to increase the yields of a cytoplasmic fusion protein [42•].

Structural disulfide bonds do not form in the cytoplasm of wild-type E. coli strains, as this environment is reducing and at least five proteins (thioredoxins 1 and 2, and glutaredoxins 1, 2 and 3, the products of the trxA, trxB, trxC, and grxE, genes, respectively) are involved in the reduction of disulfide bridges that transiently arise in cytoplasmic enzymes [43**]. Nevertheless, disulfide-bonded recombinant proteins can accumulate in the cytoplasm of surprisingly healthy trxB mutants that lack thioredoxin reductase, a protein responsible for the reduction of oxidized thioredoxins. Oxidation occurs post-translationally and is favored at low temperatures (see [44] and references therein). While mutants lacking both trxA and genes involved in glutaredoxins (e.g. grxA and grxB) are even more efficient at accumulating oxidized recombinant proteins in dithiothreitol-free medium, they exhibit severe growth deficiencies in the absence of the reducing agent [45]. As the majority of a cysteine-rich eukaryotic protein was found to accumulate in an almost completely oxidized, but inter-molecular disulfide bonded form in trxB mutants held on ice [44], the main challenge will be to engineer protein disulfide isomerases capable of reshuffling disulfide bridges into their native pattern in this environment.

**Cytoplasmic degradation**

Protein folding and proteolytic degradation are intimately linked as catalysis is an efficient way to conserve cellular resources by recycling improperly folded or irreversibly damaged proteins into their constituent amino acids. In the cytoplasm of E. coli, most — if not all — early degradation steps are carried out by five ATP-dependent Hsp70s: Lon/La, FtsH/HflB, ClpAP, ClpXP, and ClpYQ (Hu/Un) [46]. ClpAP and ClpXP act as two-component proteases that share the same degradation subunit (ClpP) but have different ATPase regulatory subunits (ClpA or ClpX). The latter appear to bind substrates in a chaperone-like manner and use ATP hydrolysis to feed them to the proteolytic center of mini-proteasome structures. Along with FtsH (an inner membrane-associated protease the active site of which faces the cytoplasm), ClpAP and ClpXP are responsible for the degradation of proteins modified at their carboxyl termini by addition of the non-polar destabilizing tail AANDENTIALAA (using amino acid single letter code) [47•,48•]. The tagging mechanism involves the 16S (i.e. stable RNA and is designed to prevent ribosome stalling at the 3′ end of damaged mRNAs [49]). Proteases Lon and ClpYQ appear to be more generic as they efficiently degrade puromycin-truncated proteins; however, there is some evidence that they also exhibit tail specificity. An obvious consequence of the existence of the SsrA tagging system is that any heterologous proteins rich in non-polar residues at its carboxyl terminus will be an appetizing substrate for cellular proteases.

A possible strategy to avoid degradation is to make use of host strains bearing mutations in protease genes; however, there are drawbacks to this approach. For example, inactivation of Lon leads to filamentation and FtsH is an essential protein for which only thermosensitive mutants are available. In addition, several proteases are usually involved in the degradation of a given protein substrate but multiple mutations in genes encoding proteases reduce cell growth rates and compromise strain fitness. An alternative is to target the polypeptide of interest to the insoluble fraction of the cell, as inclusion-body proteins are generally protected from degradation. For a normally soluble protein, this can be achieved by using strains bearing thermosensitive mutations in the major molecular chaperone systems [50]. It is important, however, to bear in mind that certain proteases (e.g. OmpT) adsorb to the surface of inclusion bodies during the recovery process and may degrade the desired protein while it is being refolded. The inner membrane protease FtsH is also active under denaturing conditions and can process recombinant proteins associated with the inner-membrane during their refolding (KW Cooper, F Baneyx, unpublished data).

**Fusion proteins**

Although fusion proteins were originally constructed to facilitate protein purification and immobilization and to couple the activity of enzymes acting in a single metabolic pathway, it soon became apparent that certain fusion partners could greatly improve the solubility of passenger proteins that would otherwise accumulate within cytoplasmic degradation systems. Suitable for the construction of fusions to maltose-binding protein (MBP), thioredoxin and glutathione S-transferase are commercially available and additionally ‘solubilizing’ fusion partners (e.g. variants of DsbA and gpHD) have recently been described [42•,51•]. The most probable reason for improved folding and/or reduced degradation of passenger proteins is that the fusion partner efficiently and rapidly reaches a native conformation as it emerges from the ribosome (or soon after its release), and promotes the acquisition of correct structure in downstream folding units by favoring on-pathway isomerization reactions. In the case of unfused cytoplasmic MBP, proper folding requires both DnaK-DnaJ-GrpE and GroEL-GroES, which may recruit chaperones in the vicinity of the passenger protein (JG Thomas, F Baneyx, unpublished data). It has also been proposed that MBP...
may directly interact with passenger proteins [52••], thereby acting as an "intramolecular" chaperone, much like protease propeptides do [53]. These mechanisms require the MBB domain to be synthesized first and are in agreement with a study showing that, whereas mammalian asparaginyl residues are soluble when fused to the carboxyl-terminus of MBB, they become insoluble when the order of the fusion proteins is reversed [54]. It should finally be noted that, despite outlandish claims, all fusion partners are not equally proficient at alleviating inclusion body formation. In a systematic comparison of the effectiveness of various fusion partners in increasing the solubility of six aggregation-prone passenger polypeptides, Kapust and Waugh [52••] found that MBB was far superior to either thioredoxin or glutathione-S-transferase as a "solubilizing" partner.

The affinity of certain fusion partners for immobilized ligands can facilitate the purification of the desired fusion protein; however, binding usually occurs with low affinity (which precludes the use of stringent wash conditions) and can be disrupted by passenger proteins. The use of polyhistidine tags at the amino-terminus or at the junction can be disrupted by passenger proteins. The use of polyhistidine tags at the amino-terminus or at the junction region of the fusion partner can solve this problem by allowing efficient purification via immobilized metal affinity chromatography [55•,55]. An additional advantage of fusion proteins is that they appear to permit the synthesis of otherwise poorly translated polypeptides. A probable explanation for this phenomenon is that the translation of passenger proteins containing rare codons occurs with higher efficiency; however, this may also lead to Lys→Arg misincorporation at rare codons.

Currently, the main disadvantages of fusion-protein technologies are that: firstly, liberation of the passenger protein requires expensive proteases (e.g. Factor Xa and enterokinase); secondly, cleavage is rarely complete leading to reduction in yields; thirdly, additional steps may be required to obtain a pure product (e.g. formation and isomerization of disulfide bonds); and finally, solubility is never guaranteed.

Secretion

Polypeptides destined for export are synthesized as preproteins containing an amino-terminal signal sequence (leader peptide) that is cleaved during the translocation process by inner-membrane-associated leader peptidases, the active sites of which face the periplasm. Typical signal sequences are 18–30 amino acids in length and consist of two or more basic residues at the amino terminus, a central hydrophobic core of seven or more amino acids, and a hydrophilic carboxyl-terminus motif recognized by leader peptidases (usually small residues at positions –1 and –3 [A, G, or S] preceded by a helix-breaking residue at position –6 [P or G]); where +1 denotes the first amino acid of the mature protein). Many signal sequences derived from naturally occurring secretory proteins (e.g. OmpA, OmpT, PelB, β-lactamase and alkaline phosphatase) support the efficient translocation of heterologous polypeptides across the inner membrane when fused to their amino-termini. In some cases, however, preproteins are not readily exported and either become 'jammed' in the inner membrane, accumulate in precursor inclusion bodies, or are rapidly degraded within the cytoplasm. While membrane jamming is an indication that translocation may be physically impossible (e.g. in the case of large cytoplasmic proteins, unnatural fusion proteins, and mutant proteins evolved by combinatorial approaches), an improved understanding of secretory mechanisms in E. coli has provided clues to circumvent other problems.

Efficient translocation requires that secretory proteins be brought into the vicinity of the inner membrane in a loosely folded form. This is guaranteed by molecular chaperones, which can be either generic (e.g. DnaK and GroEL) or specific for secretory proteins. SecB, a tetrameric polypeptide present at low levels in the cytoplasm binds to the mature domain of a subset of preproteins destined for the outer membrane and transfers them to peripheral membrane protein SecA. The latter uses energy derived from ATP hydrolysis and the proton motive force to mediate preprotein export by cycles of insertion and de-insertion into the SecYEG translocon [56•,57•]. The signal recognition particle (SRP), which consists of a 4.5S RNA and a 48 kDa GTPase termed Ffh/P48 binds highly hydrophobic signal sequences in certain preproteins (e.g. integral inner membrane proteins) and delivers them to the peripheral membrane protein FtnY in the vicinity of SecA and SecYEG [58•]. It is therefore probable that the majority of secretory proteins are delivered to the SecA motor via a variety of targeting mechanisms for export through SecYEG; however, some inner-membrane proteins also appear to directly integrate into the lipid bilayer [59].

In view of the above mechanistic information, it is tempting to hypothesize that the misfolding and degradation of a number of heterologous proteins targeted for the periplasm results from their inefficient chaperoning to the translocase, either because they fold (or misfold) too rapidly in the cytoplasm, or because the necessary chaperone(s) are limiting. Attempts to co-overexpress SecB, DnaK-DnaJ and GroEL-GroES have met with variable success and improved secretion depends heavily on the signal-sequence-mature protein combination [60]. This suggests that the signal sequence influences secondary and tertiary structure formation in the mature region of secretory proteins, which in turn affects chaperone recognition. It may therefore be necessary to try several signal sequences and/or overproduce different chaperones to optimize the translocation of a given heterologous protein. At present, there are no reports on how overproduction of components of the SRP (and in particular Ffh) affects protein secretion. This route may be particularly valuable for improving the assembly of inner membrane proteins. It should finally be noted that strains selected for their ability to restore the export of preproteins with defective signal sequences are
and the recently discovered PpiD [67]. The primary function of a number of PPIases, including SurA, FkpA, RotA/PpiA and antibody fragments [64, 66], is to catalyze the rearrangement of disulfide bonds in cysteine-rich recombinant proteins, such as human tissue plasminogen activator [65].

In addition to inefficient secretion, one of the drawbacks of periplasmic expression is that recombinant proteins may misfold or form inclusion bodies in this cellular compartment. A systematic search for periplasmic factors improving phase display [66] led to the identification of Skp/OmpH, a protein previously implicated in the folding of outer membrane proteins [65]. In contrast to specialized periplasmic chaperones (e.g. PapCD which is involved in pilus biogenesis), Skp appears to be a broad substrate range chaperone, as its overexpression improves the folding of a number of aggregation-prone single-chain antibody fragments [64, 66]. The periplasm also contains a number of PPIases, including SurA, FkpA, RotA/PpiA and the recently discovered PpiD [67]. The primary function of SurA and PpiD is to catalyze prolyl peptide bond isomerization in outer membrane proteins. Both SurA and FkpA might, however, have a more general role as their overproduction facilitates the folding of recombinant proteins that aggregate or are degraded in the periplasm [65].

Because there is no periplasmic ATP pool, misfolded proteins are degraded by energy-independent proteases, the most active of which are DegP/HtrA and Tsp, a protease that recognizes secreted proteins tagged by the SsrA system [49]. A number of additional proteolytic enzymes are present in the periplasm and cell envelope and participate in the degradation of secreted proteins [46]. Strains lacking individual or combinations of cell envelope proteases have been constructed and, although their use can help alleviate acute degradation problems, they have the disadvantage of exhibiting slower growth rates. It should finally be noted that periplasmic proteins can be further translocated to the growth medium of the cell by deliberate permeabilization of the outer membrane using a variety of systems (see [60, 69] and references therein). This process greatly simplifies the purification of target polypeptides although it increases their dilution.

Conclusions

Recent advances in the understanding of the function, regulation and interactions of cellular gene products, together with the availability of new genetic tools, are making E. coli a more attractive host than ever for the production of heterologous proteins. The facts that only a small amount of information has been exploited for practical purposes and that many fundamental aspects of E. coli physiology remain to be uncovered will continue to fuel progress in optimizing this microorganism for protein expression. Many improvements have resulted from serendipitous discoveries (e.g. the usefulness of fusion proteins and the fact that disulfide bridges can form in the cytoplasm of E. coli strains) and this trend is likely to continue. Although certain post-translational modifications (e.g. glycosylation) will probably remain beyond the reach of E. coli, robust engineered strains suitable for the cost-effective production of a wide variety of complex eukaryotic proteins should become available in the near future.

Acknowledgements

This work was supported by awards from the Bioengineering and Environmental Systems division of the US National Science Foundation. I thank Jeff Thomas for drawing Figure 1.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:
- of special interest
- of outstanding interest


An excellent review of the mechanisms through which multiplicity plasmids (primarily ColE1) are stably maintained in E. coli.

3. Williams SG, Cranenburgh RM, Weiss AME, Wrighton CJ, Sherratt DJ: A method for construction of plasmids containing multiple copies of a gene be plasmid-borne, or that gene be operably linked to a control region. Only plasmid-containing cells survive in medium supplemented with kanamycin. The authors highlight the fact that the system may be particularly useful for producing plasmid vectors for gene therapy.


An entertaining article documenting approaches that can be used to circumvent broad patent claims protecting E. coli expression systems. A strain containing T7 RNA polymerase under lac transcriptional control is constructed by first introducing a defective version of T7 gene 1 within the chromosome (i.e. opening and recombining it through homologous recombination with the missing fragment from the same gene). A strategy for multiple integration of heterologous genes in the chromosome which does not require that the genes be plasmid-borne, or that gene and promoter be ‘operably linked’ is also described. This paper should significantly reduce claim language plagiarism and force patent lawyers to become more creative.


References and recommended reading

- of special interest
- of outstanding interest


An excellent review of the mechanisms through which multiplicity plasmids (primarily ColE1) are stably maintained in E. coli.

3. Williams SG, Cranenburgh RM, Weiss AME, Wrighton CJ, Sherratt DJ: A method for construction of plasmids containing multiple copies of a gene be plasmid-borne, or that gene be operably linked to a control region. Only plasmid-containing cells survive in medium supplemented with kanamycin. The authors highlight the fact that the system may be particularly useful for producing plasmid vectors for gene therapy.


An entertaining article documenting approaches that can be used to circumvent broad patent claims protecting E. coli expression systems. A strain containing T7 RNA polymerase under lac transcriptional control is constructed by first introducing a defective version of T7 gene 1 within the chromosome (i.e. opening and recombining it through homologous recombination with the missing fragment from the same gene). A strategy for multiple integration of heterologous genes in the chromosome which does not require that the genes be plasmid-borne, or that gene and promoter be ‘operably linked’ is also described. This paper should significantly reduce claim language plagiarism and force patent lawyers to become more creative.


References and recommended reading

- of special interest
- of outstanding interest


An excellent review of the mechanisms through which multiplicity plasmids (primarily ColE1) are stably maintained in E. coli.

3. Williams SG, Cranenburgh RM, Weiss AME, Wrighton CJ, Sherratt DJ: A method for construction of plasmids containing multiple copies of a gene be plasmid-borne, or that gene be operably linked to a control region. Only plasmid-containing cells survive in medium supplemented with kanamycin. The authors highlight the fact that the system may be particularly useful for producing plasmid vectors for gene therapy.


An entertaining article documenting approaches that can be used to circumvent broad patent claims protecting E. coli expression systems. A strain containing T7 RNA polymerase under lac transcriptional control is constructed by first introducing a defective version of T7 gene 1 within the chromosome (i.e. opening and recombining it through homologous recombination with the missing fragment from the same gene). A strategy for multiple integration of heterologous genes in the chromosome which does not require that the genes be plasmid-borne, or that gene and promoter be ‘operably linked’ is also described. This paper should significantly reduce claim language plagiarism and force patent lawyers to become more creative.


20. Ross W, Ayar SE, Selamon J, Gourse RL: Promoters with UP elements of different strengths: modular structure of bacterial promoters. J Bacteriol 1998, 179:3575-3583. This paper analyzes the effect of UP elements from nosP, cyaP, fliZ, fliA, lacI, mOf and RNA II in the activity of native and lac core promoters. Transcription enhancement in hybrid lac promoters spans a wide range of values (between 1.5 and 30-fold) and correlates well with the degree of homology to the UP consensus sequence. The authors conclude that E. coli promoters are modular composites of three RNA polymerase recognition elements - the UP sequence and the -35 and -10 hexamers - whose proximity to consensus determines promoter strength.


26. Mackie GA: Ribonuclease E is a 5′-end dependent endonuclease. Proc Natl Acad Sci USA 1999, 95:70-722. A very elegant set of experiments that establishes several key properties of Ribonuclease E. If it forms linear substrates, requires unpaired 5′ nucleotides for degradation; and cleaves RNAs with 5′ monophosphates more efficiently than those with 5′ triphosphates. These results explain how 5′ hairpin may provide their stabilizing effect.


29. Coburn GA, Mackie GA: Translational enhancement by an element downstream of the initiation codon in E. coli. J Biol Chem 1999, 274:10947-10950. Utilizing the DB of C-start as a starting point, the authors created a downstream element that perfectly matches the consensus sequence and demonstrated translational enhancement both at low temperatures and at 37°C. This study further shows that a Shine-Dalgarno sequence is required for proper DB function, and that the location of this element is feasible, as it functions equally well when positioned at either codon 1 or 6.


31. Forman MD, Stack RP, Masters PS, Hauer CB, Bacter SM: High level, context-dependent misincorporation of lysine for arginine at the 115 AGA codon. Proc Natl Acad Sci USA 1999, 96:2551-2555. In cultures grown in minimal medium and in the presence or absence of glucose, 20–25% of the yeast a1 homeodomain is found to contain misincorporated lysine at the 115 AGA codon. Misincorporation is reduced to <5% when cells are grown in LB broth. Misincorporation of codon 115 for OGC reduces misincorporation to 5% in minimal medium, while incorporation of lysine at codon 115 for OGC yields a more than 99% pure product under the same growth conditions.

32. Aiyar SE, Gourse RL, Ross W: Phaseled A-tracts of sequence (A)6TGGC(A)5TAGGC)n are shown to improve yields. Host strains bearing a low-temperature inducible P1 UP element and decreas-


This paper examines the effect of co-overexpressing various thiol/disulfide type plasminogen activator in wild-type cells. Not only to facilitate the secretion of hard to translocate proteins but also to 'proofreading' function. As a result, authentic secretory proteins increases in completely absent) signal sequences. It is also observed that translocation of SecA–precursor complexes remain bound at the translocon for longer periods of time, which permits the secretion of proteins with defective (or completely) signal sequence and the major chaperone proteins on the export of human cytoxins in E. coli. A systematic search for factors improving the phage display of a fusion body is completely abolished.


Crosslinking experiments show that upon release from the SRP–Phlp complex, precursor proteins are found exclusively in the vicinity of SecY and SecE. The important result suggests that the multiple pathways for protein targeting (e.g. Secol, SRP, and proteasomal/chaperone-dependent pathways) all employ SecY and the SecX/EG translocase to mediate protein translocation across the cytoplasmic membranes.


A long overdue explanation of why ptdol mutations (which map in the sec+ gene) are such excellent suppressors of signal sequence mutations. Studies with inverted membrane vesicles (IMVs) show that SecY has a much higher affinity for the mutant form of SecY than for the wild-type version and that SecY–preprotein complexes remain bound at the translocon for longer periods of time, which again allows for the secretion of precursor proteins in ptdol-IMVs due to the lack of SecY ‘proofreading’ function. As a result, ptdol mutants should be powerful hosts not only to facilitate the secretion of hard to translocate proteins but also to further improve that of proteins which efficiently reach the periplasm in wild-type cells.


This paper examines the effect of co-overexpressing various thiol/disulfide oxidoreductases in the E. coli periplasm on the recovery of active human tissue plasminogen activator (tPA, a protein that contains 17 disulfide bridges in its native form). The folding bionner tested include E. coli DsbA (an soluble periplasmic protein that primarily functions as a strong oxidizer), E. coli DsbC (a soluble oxidase), and the three periplasmic P450s (P450, and two eukaryotic P450s from rat and yeast). Co-overexpression of DsbA is found to greatly increase tPA recovery in both shake flasks and fermentors and the purified protein has the same specific activity and sensitivity decrease in liquid cultures. In this paper, BRP mutants that retain their ability to release protein into the growth medium but do not have a detrimental effect on cell growth are selected and characterized.

64. Borth R, Plüschke A: Selection for a periplasmic factor that improves the phase display of a fusion between a poorly folding single-chain antibody fragment (SfFv) and the gene-3 protein is conducted by selecting a genomic library on the phagemid. The only event that favors display encodes Skp, a protein involved in outer membrane protein biogenesis. Skp co-overexpression is also shown to improve the folding of numerous SfFvS secreted in the periplasm. These important results suggest that Skp acts as a general periplasmic molecular chaperone. However, the authors caution that the effect of Skp may be indirect and related to the ability of this protein to prevent deleterious interactions between recombinant proteins and topoisomerase II by improving the transport of the latter species.


This study shows that Skp coexpression can greatly improve the solubility of a highly toxic SfFv secreted in the periplasm and allow it a phase display. Skp accumulation is higher when the protein is synthesized from a eukaryotic operon than when provided on compatible plasmids under an transcriptional control. High levels of Skp correlate with increase SfFv solubility.


A new heat shock gene, ppiD, which encodes a peptidyl-prolyl isomerase, is shown to be involved in the maturation of outer membrane proteins and double mutants in ppiD and surf are found to be lethal. By contrasting combinations of ppiD mutations with deletions in other periplasmic PPiases (ppiA-ppiB) or skp do not confer a lethal phenotype.


Bacteriocin-release protein (BRP) is a small lipoprotein that causes the release of periplasmic proteins into the growth medium when secreted into the cell envelope. Its usefulness is dampened by the fact that, even when secreted from the Lpp leader peptide, accumulation of mature BRP causes toxicity decrease in liquid cultures. In this paper, BRP-mutants that retain their ability to release proteins into the growth medium but do not have a detrimental effect on cell growth are selected and characterized.


TolAIII co-overexpression in promoting the efficient release of secreted and periplasmic protein products is demonstrated. It is noted that although the strategy reduces the total levels of recombinant (Bacto) mass secreted from the ompA- leader, formation of periplasmic inclusion bodies is completely abolished.


The only insert that favors display encodes Skp, a protein involved in outer membrane protein biogenesis. Skp co-overexpression is also shown to improve the folding of numerous SfFvS secreted in the periplasm. These important results suggest that Skp acts as a general periplasmic molecular chaperone. However, the authors caution that the effect of Skp may be indirect and related to the ability of this protein to prevent deleterious interactions between recombinant proteins and topoisomerase II by improving the transport of the latter species.