*Essays in Biochemistry* (2021) **65** 247–260
https://doi.org/10.1042/EBC20200170

PORTLAND PRESS

Review Article

# So you want to express your protein in *Escherichia coli*?

**Aatir A. Tungekar*, Angel Castillo-Corujo* and Lloyd W. Ruddock**

Protein and Structural Biology Research Unit, Faculty of Biochemistry and Molecular Medicine, University of Oulu, Oulu 90220, Finland

**Correspondence:** Lloyd W. Ruddock (lloyd.ruddock@oulu.fi)



Recombinant proteins have been extensively employed as therapeutics for the treatment of various critical and life-threatening diseases and as industrial enzymes in high-value industrial processes. Advances in genetic engineering and synthetic biology have broadened the horizon of heterologous protein production using multiple expression platforms. Selection of a suitable expression system depends on a variety of factors ranging from the physicochemical properties of the target protein to economic considerations. For more than 40 years, *Escherichia coli* has been an established organism of choice for protein production. This review aims to provide a stepwise approach for any researcher embarking on the journey of recombinant protein production in *E. coli*. We present an overview of the challenges associated with heterologous protein expression, fundamental considerations connected to the protein of interest (POI) and designing expression constructs, as well as insights into recently developed technologies that have contributed to this ever-growing field.

## Introduction

Ever since the Food and Drug Administration (FDA) approved the first recombinant protein for therapeutic use in 1982, *Escherichia coli* has been a workhorse for recombinant protein production in both academia and industry. Despite huge advances in other expression systems, the production of heterologous recombinant proteins in microbial expression systems remains simpler and less expensive than in alternative systems such as mammalian cell culture [1]. *E. coli* offers various advantages such as comparatively easier genetic manipulation, use of simple growth medium, rapid cell growth, simple fermentation process, virus-free product, high product yields, and cost-effective production [1]. The science behind recombinant protein production seems straightforward, however, in practice, multiple factors can impose hurdles. As Sun Tzu says in the Art of War '*know the enemy and know yourself*', because if you do not then there is a high chance of failure. Hence, the starting point for any expression should be to know your protein.

## The protein and its properties

This review will focus on the production of soluble proteins or soluble fragments of transmembrane (TM) or membrane-associated proteins. For additional issues connected with the production of TM proteins, see [2–4]. Often the protein of interest (POI) is a eukaryotic protein. This can cause additional problems including codon usage, post-translational modifications (PTMs) and issues related to protein folding.

For an overview of the full workflow, see Figure 1. The starting point for any protein expression is to define the protein you wish to make, taking into account possible splice variants, signal sequences, TM helices, and PTMs found in the natural protein. While protein databases such as UniProt [5] are an excellent starting point for looking at these, it is always worthwhile doing additional bioinformatics analysis (Table 1).

While bioinformatics approaches are powerful, they are only predictions and so gathering a consensus from multiple independent bioinformatics approaches or looking for validation through experimental

*These authors contributed equally to this work.

Received: 28 February 2021
Revised: 27 March 2021
Accepted: 30 March 2021

Version of Record published: 26 July 2021

*Essays in Biochemistry* (2021) **65** 247–260
https://doi.org/10.1042/EBC20200170

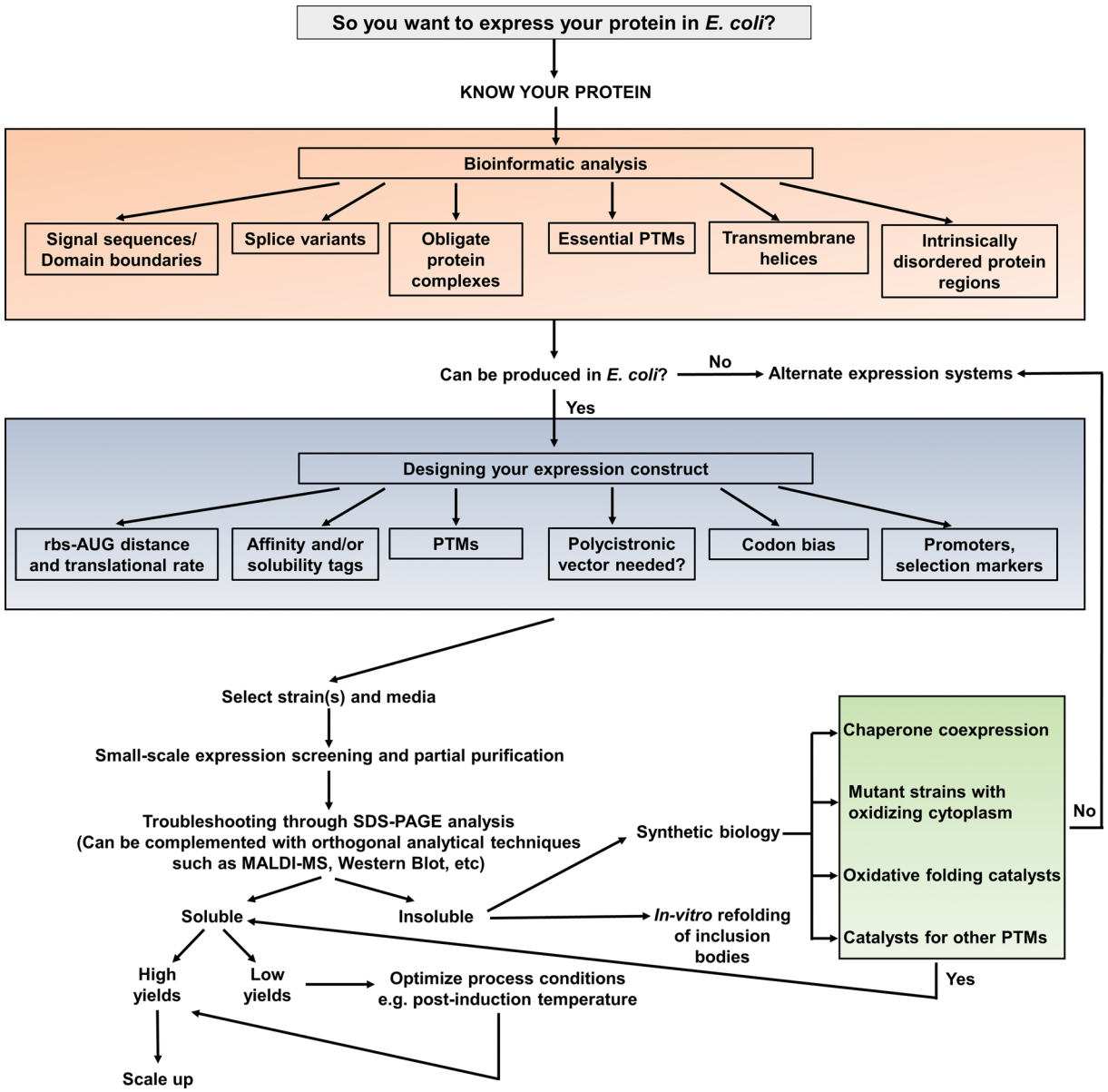


**Figure 1. Overall workflow**

Schematic showing a simplified workflow. In practice, it may be more iterative than this.

means (e.g., from published literature) is always worthwhile. For example, human cytotoxic T-lymphocyte antigen 4 (CTLA-4) is an obligate dimer and requires N-glycosylation of $Asn^{78}$ and $Asn^{110}$ for dimerization [6]. As this PTM cannot be made in *E. coli*, spending a little time to know your protein can save a lot of heartache later on. In essence, without the use of synthetic biology approaches (see below), the only eukaryotic-like PTMs *E. coli* does is disulfide bond formation in the periplasm [7].

It is also often worthwhile using bioinformatics approaches, e.g. JPRED [8] to look for both domain boundaries and prediction of intrinsically disordered protein (IDP) regions. Expressing a construct that is too short and misses an essential part of a domain, e.g. a β-strand, is always going to result in failure, while expressing a construct that is too long and includes flexible regions prone to proteolysis is likely to either result in heterogeneity or loss of a purification tag. Proteins with large IDP regions are often problematic to make as they are often prone to degradation, however, it should be remembered that many IDP regions may gain structure upon interaction with other molecules, e.g. upon protein complex formation (e.g. ACTR and nuclear co-activator binding domain (NCBD)) [9] and so, co-expression of a partner may help considerably in obtaining the protein in a stable and soluble form.

**Table 1 Suggested bioinformatics analyses to be undertaken before starting to clone the gene for your protein**

| Bioinformatics analysis | Examples | Comments |
|---|---|---|
| Signal sequences | SignalP [70] | Should not be included in the protein sequence you want to express in *E. coli*. If you want to target the protein to the periplasm, use a cleavable *E. coli* signal sequence. See [71] for a recent review. |
| TM helices | TOPCONS [72] | TM helices should not be included in the protein sequence if you want to obtain soluble protein. |
| Glycosylation | Reviewed in [73] | Our rule of thumb is that if the protein has more than one N-glycosylation site per 100 amino acids the protein it may not be expressed solubly, as glycans enhance solubility and *E. coli* does not naturally glycosylate proteins. |
| Disulfide bonds | UniProt or scientific literature | Most proteins that contain structural disulfides require them to be natively formed to allow soluble protein production. In our experience, disulfide bond prediction can be poor except by homology. A good rule of thumb is that proteins that enter the secretory pathway that contains cysteines are likely to contain disulfide bonds. |
| Other PTMs, e.g. Sulfation phosphorylation | Sulfinator [74] NetPhos [75] | Our rule of thumb is that while these may modulate the function of the protein, their absence does not affect soluble protein production. |
| Biophysical properties | Protparam [76] | The pI and molecular weight of the protein are useful for confirming expression and for rational protein purification, e.g. the calculated pI helps predict column types and pH for ion-exchange chromatography. |
| Complex formation | UniProt or scientific literature | Obligate protein complexes usually require most/all the proteins in the complex to be co-expressed to be able to obtain folded soluble protein. |

For other analyses, see for example ExPASy [66]. Abbreviation: pI, isoelectric point.

Before cloning the gene for the protein you want, it is worth considering how you are going to subsequently purify it, as this may affect the construct you want to express. The most powerful first step in the purification of soluble proteins is affinity chromatography (if possible). This includes either the endogenous properties of the protein, e.g. immobilized-ligand or substrate mimic chromatography (e.g. Cibacron Blue F3GA [10] or cyclic peptide-based ligands [11]) or the addition of a tag to aid purification, e.g. a maltose-binding protein (MBP)-tag, glutathione-S-transferase (GST)-tag or most commonly a hexahistidine tag (His-tag) allowing the use of immobilized metal affinity chromatography (IMAC). For an overview of possible affinity tags, refer to [12]. If the structure of your protein or something closely related is available, it is worthwhile looking at the accessibility of the N- and C-termini to see if any added tag is likely to be disruptive to the structure, e.g. if the protein termini are buried. Alternatively, structure prediction programs such as Phyre 2 [13] could be used. While very useful and widely used, N-terminal His-tags may increase the heterogeneity of your final product due to variable (phospho)gluconylation of the N-terminus [14].

Depending on the end use of the protein, you may want to be able to remove the affinity tag after purification by proteolysis. Enzymes with broad specificity can sometimes be used, e.g. trypsin can be used to both remove an N-terminal tag and the C-peptide from insulin derivatives, e.g. [15] but usually, removal of affinity tags is mediated through more highly specific proteases such as TEV (consensus site ENLYFQ↓G/S) and Factor Xa (consensus site IE/DGR) [12]. Care should be taken of the source of the protease, for example, recombinant bovine Factor Xa is reported to have a different specificity than recombinant human Factor Xa [16,17]; see also MEROPS database for other proteases [18]. Most proteases have specificity to sequences both before and after the site of cleavage and so often one or more amino acids from the cleavage site are left on the mature protein. In addition, proteases cannot access buried cleavage sites and so often the cleavage site is put into a flexible linker region (usually glycine/serine-rich), which may add more residues to the mature protein.

In addition to making fusion proteins to aid purification, they can also be used to add solubilization tags. Such tags which are often small, highly soluble, and stable proteins, can aid not only in the solubilization of the final product but also in the solubilization of folding intermediates. If a eukaryotic protein has more than one N-glycan per 100 amino acids, a solubilization tag may be essential to produce it in a soluble form in *E. coli*. Commonly used solubilization tags include MBP (which doubles as an affinity purification tag), thioredoxin, Sumo, or Fh8. For solubilization tags, there needs to be a balance, if they help too little then soluble protein may not be achieved. Conversely, if they help solubilize too much then false positives may be achieved where the final product is soluble despite the POI not being correctly folded. This balance often has to be achieved by trial and error.

Even with careful selection of domain boundaries and possible solubilization tags, not all eukaryotic proteins can fold to a native state in *E. coli*. This is linked to issues of protein folding, PTMs, and/or the protein being part of an unknown obligate complex. *E. coli* contains a wide range of molecular chaperones (e.g. GroEL/ES, DnaK, Skp) and ten

*Essays in Biochemistry* (2021) **65** 247–260
https://doi.org/10.1042/EBC20200170

peptidyl *cis-trans* prolyl isomerases and so issues related to protein folding are usually either linked with (i) translation rates (see below); (ii) oxidative folding, i.e. the formation of disulfide bonds; (iii) the protein having an essential PTM which *E. coli* cannot perform; (iv) the protein having a buried prosthetic group which wildtype *E. coli* cannot make or becomes limiting (in some cases this can be solved by the addition of the moiety to the growth media); (v) rare cases where a specialized folding factor is involved in folding the protein, e.g. to express a hyperthermophilic α-amylase from *Pyrococcus furiosus* (a hyperthermophilic archaeum) in *E. coli*, the co-expression of small heat shock protein (sHSP) or chaperonin (HSP60) from the same *P. furiosus* was found to be essential [19]. For an overview of alternate expression platforms and genetic engineering approaches available to carry out PTMs in heterologous proteins, refer to [20].

Native disulfide bond formation is the most common issue. There are three approaches to deal with this issue. Firstly, the protein could be allowed to form aggregates, or inclusion bodies, of misfolded/unfolded protein. Inclusion bodies are relatively easy to purify, and the protein can then be refolded *in vitro* [21,22]. Secondly, the protein could be targeted to the periplasm via the addition of an N-terminal periplasmic signal sequence. Here there is machinery for native disulfide formation [7], and while it is a powerful technique both the sec secretion system and the folding apparatus in the periplasm can easily be overwhelmed, so (extreme) care must be taken [23]. Thirdly, an engineered strain could be used that removes disulfide bond reducing pathways from the cytoplasm [24,25], or adds oxidative folding catalysts, reviewed in [26]. This can be combined with the TAT-secretion system for exporting folded proteins to the periplasm, e.g. [27,28]. Similar synthetic biology approaches also allow other PTMs to be made in the cytoplasm, for example mucin-type O-glycosylation in *E. coli*. [29].

Finally, it should be remembered that the cytoplasm of *E. coli* contains methionine aminopeptidase, which can remove the initiating methionine [30], depending on the subsequent amino acids (e.g. serine, alanine, cysteine, proline, or glycine at P1′ preferred, Pro at P2′ inhibits), with engineered systems extending the list, e.g. [31]. This also combines with the N-end rule for protein clearance from a cell. For *E. coli*, proteins with an N-terminal Arg, Lys, Leu, Phe, Tyr, or Trp can be rapidly degraded [32], but this depends on the context of the N-terminal and subsequent amino acids [33,34].

After all these considerations, if no purified protein is obtained, a simple troubleshooting sodium dodecyl sulfate/polyacrylamide gel electrophoresis (SDS/PAGE) analysis may quickly help elucidate the possible issues (Figure 2). SDS/PAGE analysis can be complemented by other techniques including mass spectrometry, Western blotting, activity assays for the POI etc.

## The gene and its properties

Once details of the protein construct are finalized it is time to turn your attention to the gene. Just as much care must be taken for it as for the protein construct or yields may be low. One important concept that is often forgotten in protein expression is cellular homeostasis or everything in balance. Too often a high-copy number plasmid may be used with a strong promoter, but this will invariably result in less protein than could be produced as too many cellular resources are put into making plasmid deoxyribonucleic acid (DNA) and messenger RNA (mRNA), and the mRNA produced is in far excess of the limitations of the translation apparatus (Figure 3).

A multitude of genetic engineering strategies have been developed over the years to enable efficient cloning of protein expression constructs [35,36]. While industry often integrates genes into the bacterial chromosome to avoid the problem of plasmid loss during large scale fermentation, the academic approach more usually uses plasmids for expression as they are faster and cheaper to use. Plasmid selection for protein production is based on (i) copy number, which depends on the origin of replication of the plasmid (Table 2); (ii) promoter (Table 3); (iii) selection marker (Table 4). There is a balance between plasmid copy number and promoter strength (Figure 3) to maximize cellular resources going into protein production and this also depends on the media, with chemically defined minimal media being more sensitive to alterations in these, in particular when either is excessively high. Recent advancements in synthetic biology led to growth-decoupled recombinant protein production through the co-expression of a bacteriophage-derived *E. coli* ribonucleic acid (RNA) polymerase inhibitor peptide called Gp2 [37]. This approach allowed the modulation of metabolic resources, so they are exclusively utilized to produce the POI.

The plasmid is not the only decision to make. The source of the gene is important. For decades, the normal source of the gene for the POI was directly from the original organism e.g., by complementary DNA (cDNA) library obtained by real time-polymerase chain reaction (RT-PCR) from an mRNA pool (to avoid introns). While this can be fast, cheap and efficient, it can give rise to problems connected with differences in translation initiation and codon usage between prokaryotes and eukaryotes.

**Table 2 Origins of replication and associated copy numbers of vectors used for protein production in *E. coli***

| Ori | Typical copy number | Example vector | References |
|---|---|---|---|
| pMB1 | ∼15–20 | pBR322 | [77] |
| pMB1 (derivative) | ∼500–700 | pUC/pGEM | [78] |
| pBR322 | ∼15–20 | pET/pGEX | [79] |
| ColE1 | ∼15–25 | pColE1 | [80] |
| ColE1 (derivative) | ∼300–500 | pBluescript | [81] |
| p15A | ∼10 | pACYC | [82] |
| R6K | ∼10–15 | pR6K | [83] |
| pSC101 | ∼5 | pSC101 | [84,85] |

SC101 or R6K are compatible with p15A and with one from the set of pMB1/pBR322/ColE1, i.e. they can exist in the same cell on different plasmids while other combinations are not compatible. R6K and pSC101 require *pir* and *dnaA* genes respectively, for replication.

**Table 3 Common promoters used by academia and industry for recombinant protein production**

| Promoter | Comments |
|---|---|
| Lac | Relatively low constitutive expression in the absence of the lacI repressor. Inducible by IPTG and allolactose (formed from lactose by the action of LacZ. Repressed by glucose. |
| LacUV5 | Similar to the lac promoter but stronger due to more efficient recruitment of RNA polymerase. |
| T7 | Based on T7 bacteriophage system which promotes high levels of transcription. This promoter cannot be recognized by the host polymerase, so requires T7 polymerase—often chromosomally integrated under the control a LacUV5 promoter. |
| T5 | Based on T5 bacteriophage early promoter and the lac-operon. It contains three LacI binding sites and is strongly repressed in $LacI^q$ strains. Inducible by IPTG and lactose. |
| Tac | A developed hybrid of lacUV5 and trp promoters. Higher expression levels than either, with tight regulation. Inducible by IPTG and allolactose and repression by LacI and glucose. |
| araBAD | Tunable induction by L-arabinose. Tightly regulated independent of the presence of other carbon sources. Depends on Ara status of the host cell. |
| rhaBAD | Tunable induction by L-rhamnose. Low basal expression. Tightly regulated independent of the presence of other carbon sources. |
| proU | Promoter from an osmoregulated operon. Induction by higher osmolarity, e.g. increased [NaCl] in the media. |

The table includes details of promoters ranging from operons induced by sugar substrates to those induced by the ionic strength of the media (reviewed in [67,68]). Abbreviation: IPTG, isopropyl β-D-1-thiogalactopyranoside.

**Table 4 Antibiotic- and non-antibiotic-based selection markers used for clone selection and plasmid maintenance in recombinant protein production**

**Antibiotic based**

| Name | Mechanism of action and inactivation | References |
|---|---|---|
| Ampicillin | Acts as an inhibitor of transpeptidase and causes cell lysis. The $amp^r$ gene encodes β-lactamase which catalyzes the hydrolysis of the B-lactam ring of ampicillin. | [86] |
| Chloramphenicol | Acts to inhibit protein synthesis by the ribosome and hence is bacteriostatic. The $cam^r$ gene encodes an acetyltransferase that, catalyzes the formation of inactive hydroxyl acetoxy derivatives. | [87] |
| Kanamycin | Binds to 30S ribosome subunit and causes misreading of mRNA. The $kan^r$ gene encodes for an enzyme that phosphorylates kanamycin, thereby inactivating it. | [88] |
| Tetracycline | Tetracycline blocks the A site of the ribosome preventing entry by tRNAs. The $tet^r$ gene encodes an efflux protein transporting tetracycline out of the cytosol. | [89] |
| Streptomycin | Streptomycin binds to 16S ribosomal RNA, inhibiting protein synthesis. The $strep^r$ gene encodes for aminoglycoside modifying enzymes such as nucleotidyltransferases or phosphotransferases which inactivate streptomycin. | [90] |
| FabV-Triclosan | Plasmid system expressing FabV which protects against deleterious effects of Triclosan added to the growth media. | [91] |

**Non-antibiotic based**

| Name | Mechanism | References |
|---|---|---|
| Gene KO | Plasmid carries the wildtype gene to complement the auxotrophy in a knockout *E. coli* strain, e.g. Δ*ProBA, ΔTpiA, ΔglyA*, Δ*QAPRTase*. | [92–95] |
| *lac-DapD* | Plasmid-mediated repressor titration: The engineered host strain contains *dapD* under control of the *lac* operator/promoter *(lacO/P). A plasmid containing lacO releases repression of DapD by titration of lacI.* | [96] |
| ColE3-Amn | The vector contains the C-terminal ribonuclease domain of colicin E3 with an amber stop codon (s) at 5′ terminus. Allows propagation of the vector in *E. coli* cells without amber suppressor activity. | [97] |

*Essays in Biochemistry* (2021) **65** 247–260
https://doi.org/10.1042/EBC20200170

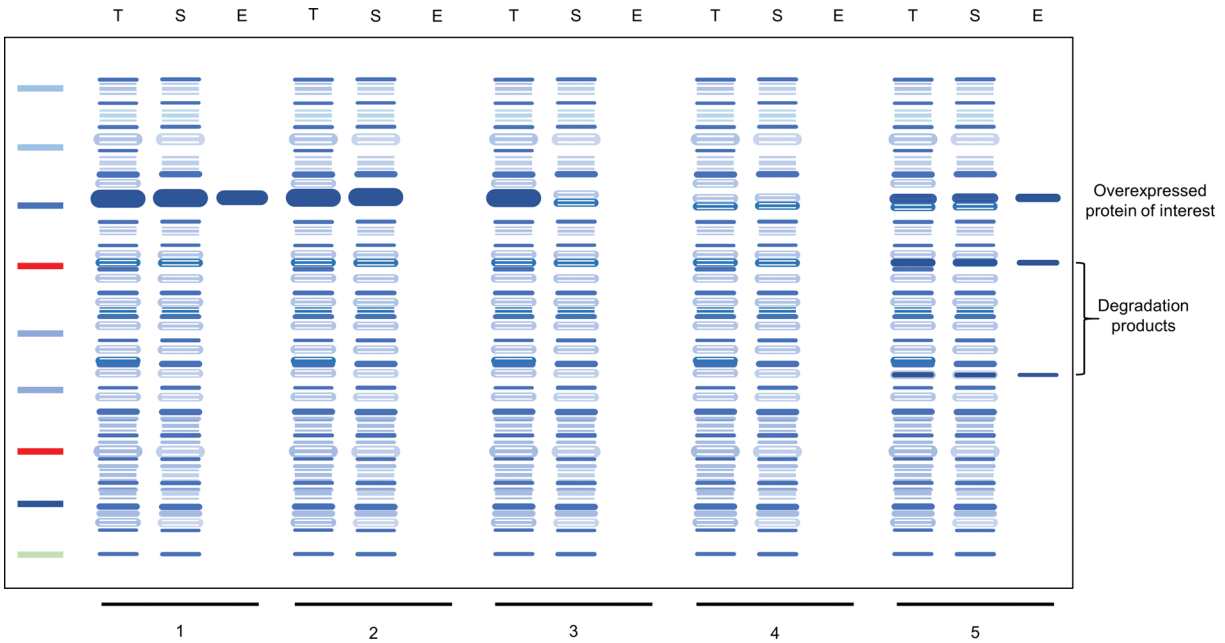


**Figure 2. SDS/PAGE gel troubleshooting**

Schematic of how SDS/PAGE analysis of total cell lysate (T), soluble fraction of the lysate (S) and purified protein (E), e.g. from IMAC of a His-tagged protein, can help to narrow down the cause of problems with production of a POI. **1**: Everything goes well. The soluble expression level is equal to the total expression level and the protein can be purified. **2**: The POI is expressed solubly, but cannot be purified, e.g. due to accessibility or proteolytic removal of the purification tag. **3**: The protein band is only visible in the total lysate lane indicating no soluble protein was made, due to either folding issues or the presence of a membrane associating region. **4**: The absence of visible POI indicates expression issues, e.g. incorrect induction or no translation initiation or very high sensitivity to proteolysis. **5**: The POI is expressed and soluble, but susceptible to proteolytic degradation.

While eukaryotic ribosomes bind to the cap at the 5′ end of the mRNA and then move down the mRNA until they initiate translation from the first AUG codon with a Kozak sequence in front of it, prokaryotic ribosomes bind to a sequence on the mRNA known as the Shine–Dalgarno (SD) sequence or ribosome-binding site (rbs; Figure 4). The rbs are usually 5–13 base pairs [38] upstream of the initiating AUG (optimal distance 5–6 base pairs [39]); and are complementary to the 3′ end of the 16S ribosomal RNA. In *E. coli*, this sequence is AGGAGGU [40]. The requirement for a distinct rbs has two consequences for eukaryotic protein expression in *E. coli*. Firstly, an rbs must be present before the initiating AUG. This may be present in the plasmid outside the multicloning site, but care should be taken that it is within the correct distance and that there are no other possible AUG trinucleotides that translation could initiate from. Secondly, this nucleotide sequence should not appear inside the gene of interest. An internal rbs will either result in the generation of a second protein (if there is an AUG at the correct distance from it) or will result in translation stalling as a ribosome binds to this site and prevents translation through it. Due to this care must be taken in the codon used for Gly–Gly pairs (i.e. not GGA–GGU), Arg–Arg pairs (i.e. not AGG–AGG), and sequences around Glu (GAG), including Glu–Glu pairs (GAG–GAG). AGG and GGA codons are rarely used by *E. coli* (see below) and so mostly care with codon optimization to avoid internal rbs relates to sequences around Glu (Q/K/E-E or E-V).

Codon usage is not equally distributed among the codons available and the variation in codon usage bias is considerable between organisms (Table 5). Codon usage varies considerably between organisms (Table 5) and correlates with corresponding transfer RNA (tRNA) levels [41]. mRNA which contains multiple rare codons can exhibit translation stalling and mRNA degradation, reviewed in [42]. Codon usage issues can be examined by bioinformatic approaches, e.g. Graphical Codon Usage Analyzer [43]. One method to prevent this problem was the overexpression of rare tRNAs, e.g. [44,45] such as from pLysSRARE [46]. For more detailed insights into codon usage, refer to [47]. The more usual approach now is the use of synthetic genes that can be codon optimized for the expression host, while simultaneously avoiding internal rbs, internal restriction sites, and factors that influence mRNA structure and stability [48,49]. As

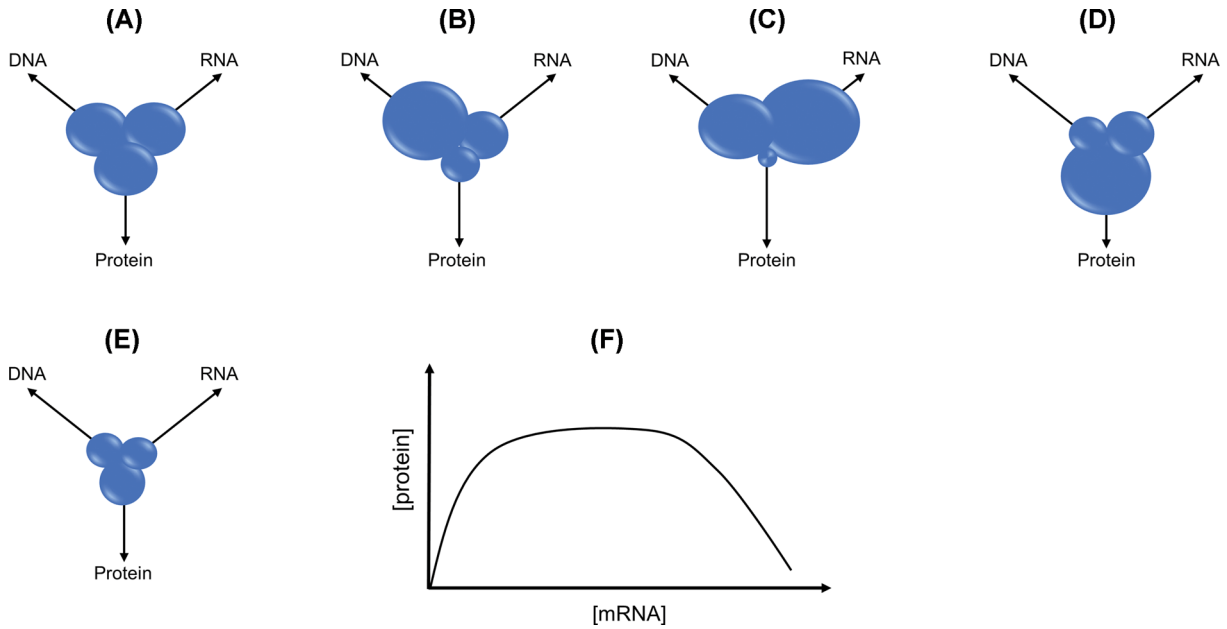


**Figure 3. Proteostasis and the balance between gene copy number, promoter strength and recombinant protein expression levels**

(**A**) Cellular resources are split evenly between DNA replication, RNA, and protein production (as well as other cellular processes not shown). (**B**) Too high plasmid copy number increases cellular resources needed for DNA replication, limiting those available for other processes including recombinant protein production. (**C**) A high copy number and a highly induced strong promoter can significantly reduce protein production. (**D**) Having the plasmid copy number and promoter strength just right can lead to maximal protein production. (**E**) Even with the optimal copy number and promoter strength, too low induction results in lower than optimal protein production. (**F**) Soluble protein yields increase with [mRNA], plateau, and then decrease as either too many resources go into mRNA production and/or the protein folding capacity of the cell is overloaded.

**Table 5 Codon usage bias ranked by *E. coli* usage**

| Codon | Amino acid | Expected usage | *E. coli* (W3110) | | *Homo sapiens* | |
|---|---|---|---|---|---|---|
| | | | Usage | Ratio | Usage | Ratio |
| AGG | Arg | 0.17 | 0.02 | 7.2 | 0.21 | 0.8 |
| CUA | Leu | 0.17 | 0.04 | 4.7 | 0.07 | 2.4 |
| AGA | Arg | 0.17 | 0.04 | 4.0 | 0.22 | 0.8 |
| UAG | Stop | 0.33 | 0.06 | 5.2 | 0.24 | 1.4 |
| AUA | Ile | 0.25 | 0.07 | 3.6 | 0.17 | 1.5 |
| CUC | Leu | 0.17 | 0.10 | 1.6 | 0.20 | 0.8 |
| GGA | Gly | 0.25 | 0.11 | 2.3 | 0.25 | 1.0 |
| CCC | Pro | 0.25 | 0.12 | 2.0 | 0.32 | 0.8 |
| ACA | Thr | 0.25 | 0.13 | 1.9 | 0.28 | 0.9 |
| GGG | Gly | 0.25 | 0.15 | 1.7 | 0.25 | 1.0 |
| CCU | Pro | 0.25 | 0.16 | 1.6 | 0.29 | 0.9 |
| GCU | Ala | 0.25 | 0.16 | 1.6 | 0.27 | 0.9 |
| AAG | Lys | 0.5 | 0.23 | 2.1 | 0.57 | 0.9 |
| GAG | Glu | 0.5 | 0.31 | 1.6 | 0.58 | 0.9 |

The expected usage for each amino acid dependence is based on the number of codons encoding that amino acid. The ratio of the expected usage to actual usage in an organism shows the relative underuse of the codon. Codon usage between *E. coli* and human genes is quite different, with only the CUA codon being relatively underused in both organisms. Codon data taken from [69].

*Essays in Biochemistry* (2021) **65** 247–260
https://doi.org/10.1042/EBC20200170

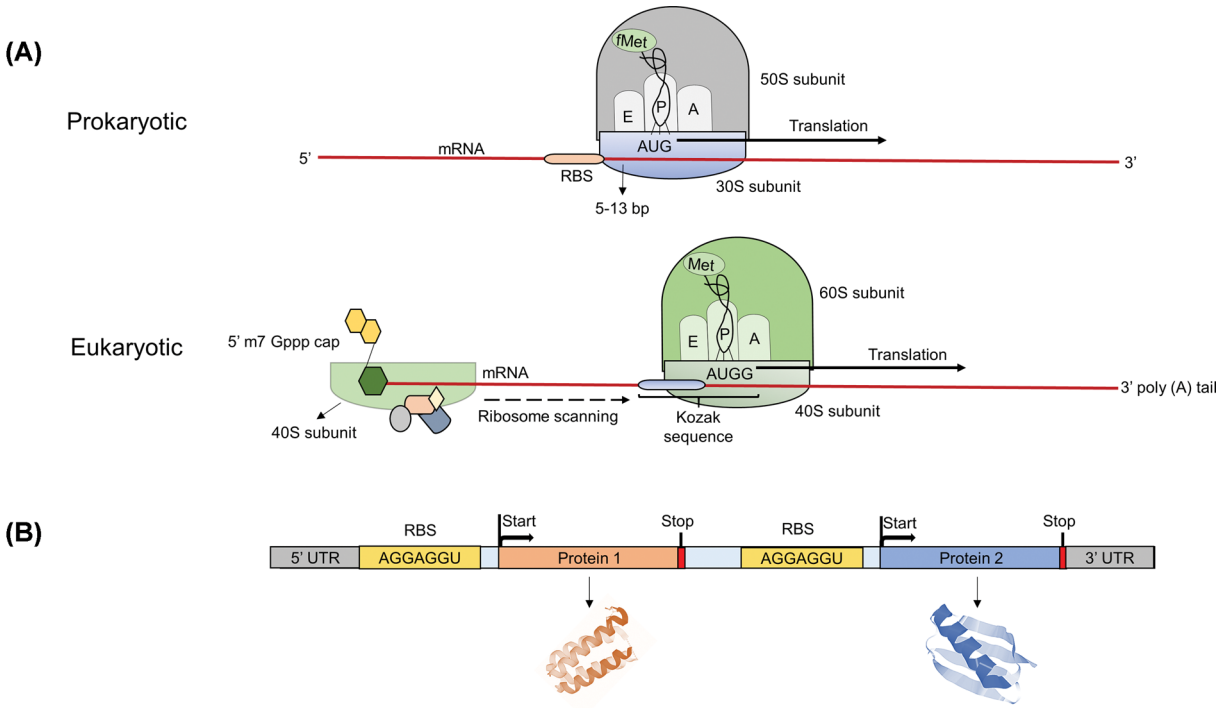


**Figure 4. Schematic representation of initiation of translation in prokaryotes and eukaryotes**

(**A**) The process of translation is carried out by a ribosome comprising the 50S (large) and 30S (small) subunits in prokaryotes and of 60S (large) and 40S (small) subunits in eukaryotes. The key difference between the two is that, in prokaryotes, the small ribosome subunit binds to the ribosome-binding site (RBS) known as SD sequence upstream of the start codon, while in eukaryotes the small ribosomal subunit binds to the 7-methylguanosine cap at the 5′ end of the mRNA. The SD sequence in prokaryotes aids in the proper aligning of the ribosome subunit to the start codon (AUG). In eukaryotes, the small ribosomal subunit bound at the 5′ end scans the mRNA in the 5′→3′ direction to locate the Kozak sequence (ACC**AUG**G) which contains the start codon. In both prokaryotes and eukaryotes, the large ribosome subunit is recruited to the mRNA once the start codon is recognized to initiate the process of translation. (**B**) Schematic representation of a polycistronic mRNA in prokaryotes. One of the key features of mRNA in prokaryotes is that they can exist in a polycistronic form, whereas the eukaryotic mRNA is monocistronic. A polycistronic mRNA consists of multiple cistrons each of which can be translated to a protein independently, i.e. a single mRNA transcript can be translated to produce more than one protein.

prices have rapidly dropped a synthetic gene can cost less than the labor and material costs associated with cloning a gene from a cDNA library.

Synthetic genes can also help mitigate the potentially deleterious effects of one other difference between eukaryotic and prokaryotic protein translation, translation rates. In prokaryotes such as *E. coli*, transcription and translation rates are coupled, with transcription rates approx. 50 nucleotides/s and translation rates approx. 16 amino acids/s [50]. In contrast, translation rates in eukaryotes are slower, with a rate of approx. 3 amino acids/s [51]. Protein folding has evolved in parallel with these translation rates and hence when a eukaryotic protein is expressed in *E. coli*, the rate of the translation may be faster than the rate of folding and for multidomain proteins, this can be a serious issue (Figure 5). This can be mitigated by modulation of translation rate [52], codon usage harmonization [53], or the use of rarer codons just after domain boundaries to cause ribosome stalling [54] (Figure 5).

A specialized ribosome system aimed specifically at the expression of the POI in *E. coli* by modifying the SD sequence of the mRNA and corresponding anti-SD sequence of the 16S rRNA was first reported by Hui and De Boer in 1987 [55]. Alternative ribosome systems such as the orthogonal riboswitch system [56], the *RiboTite* system [57], and the Ribo-T system [58] have been reported since. The riboswitch system allows tunable co-expression of multiple genes in a dose-dependent response to small synthetic molecules while the *RiboTite* system, which builds on the riboswitch technology, has been shown to harmonize protein translation rates with protein secretion [59]. The Ribo-T system employs an engineered hybrid rRNA composed of both small and large subunit rRNA sequences, in which short RNA linkers covalently link the subunits into a single translating unit [58]. This orthogonal ribosome–mRNA

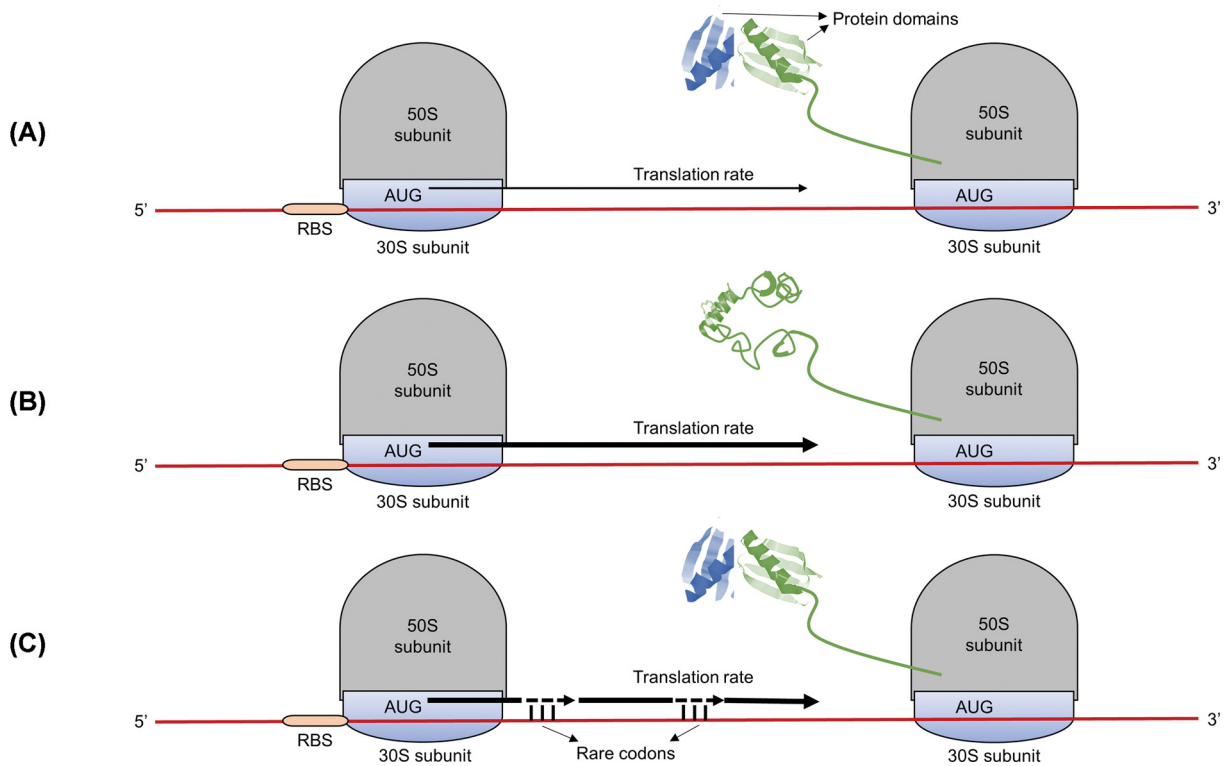


**Figure 5. The influence of translation rate on protein folding efficiency**

(**A**) When the translation rate is not too high, each domain in the protein has sufficient time to fold and the native structure is obtained. (**B**) When the translation rate is too high, individual domains of the protein might not be able to fold into their native state before the next domain is translated, resulting in inappropriate interactions between non-native domains and hence misfolding. This effect may occur when a eukaryotic protein is expressed in a prokaryotic organism due to differences in translation rates between organisms. (**C**) If rare codons are introduced at domain boundaries, the rate at which the nascent polypeptide is being translated is modulated such that the protein domains can fold and misfolding is minimized.

system is capable of supporting bacterial growth even in the absence of wildtype ribosomes and its improved tethered version has been reported recently [60].

Another difference between eukaryotic and prokaryotic protein translation can be an advantage for recombinant protein production. Many prokaryotic genes are expressed in operons, where a single promoter results in the production of multiple proteins from a single mRNA that has an rbs before the initiating AUG of each (Figure 4). This allows both the co-expression of subunits that form complexes, or the co-expression of ancillary factors that may be required for the protein to reach the native conformation.

## Strains and media for small-scale expression screening

Once a suitable construct for protein expression has been generated, the next step is to express the protein. This leads again to more rational choices needing to be made. *E. coli* is a remarkably diverse bacterial species, with only approx. 20% of the genome common to all strains [61]. It can be broadly split into four subgroupings, K-12 strains, B-strains, and the C and W strains based on their initial isolation [61]. Many K-12 and B-strains are used for recombinant protein production (Table 6). Some POI show strong strain dependence, often for unclear reasons, so we routinely test any new protein in at least one K-12 and one B-strain. Similarly, there are a wide variety of media choices, which can be broadly split into rich media (which contains yeast extract and/or another mixed source of peptides such as tryptone) and chemically defined or minimal media (where there are often only 1–3 carbon sources and a single nitrogen source). Again, some POIs show strong media dependence for production and so we routinely test any new protein in at least one rich media and one chemically defined media. While Luria–Bertani (LB) media used to be the default media for academic protein production, it has been largely superseded by media which allow higher density cultures to be obtained as higher cell mass usually results in higher protein yields. In particular, the use of

*Essays in Biochemistry* (2021) **65** 247–260
https://doi.org/10.1042/EBC20200170

**Table 6 Most common *E. coli* strains used in the production of heterologous proteins in academia and industry**

| Strain | Comments | References |
|---|---|---|
| BL21 | B-derived strain widely used for production of recombinant proteins. Deficient in lon and OmpT proteases. | [98] |
| BL21 (DE3) | Derived from BL21; routinely used for protein expression under the control of a T7 promoter regulated by T7 RNA polymerase carried by the DE3 prophage (chromosomal integration under the control of a lacUV5 promoter). | [98] |
| C41 (DE3)/C43 (DE3) | Derived from BL21(DE3) with unspotted mutations that allow them to produce some toxic and membrane proteins. | [99] |
| MG1655 | Well characterized K-12 derived strain widely used for recombinant protein production. Higher stress resistance allows high-cell density fermentation. | [100] |
| W3110 | Closely related to MG1655. Stress resilience and membrane stiffness allows high-cell density fermentations for heterologous protein production. | [100] |
| RV308 | A K-12 derived strain mutated for industrial protein production; offers increased protein yields and low acetate production. | [101] |
| HMS174(DE3) | K-12 derived strain with a recA mutation. These strains stabilize certain target genes whose products may cause the loss of the DE3 prophage and allow heterologous protein production under the control of a T7 promoter. | [101] |

For genotypes of these and other strains, see for example https://openwetware.org/wiki/E._coli_genotypes.

auto-induction media, e.g. [62], both facilitate the screening of multiple POI and allow culture densities typically 10× higher than LB. Additionally, an alternate growth medium for recombinant protein production in *E. coli* which allows the controlled release of substrates, thereby mimicking fed-batch process conditions at a small scale, has been reported [63].

In addition to strain and media, the temperature of the culture post-induction can play a key role in the yield of the folded protein. This effect probably arises both from the change in relative hydrophobicity with temperature and from the slower rate of protein translation [64] so as not to exceed the capacity of the folding machinery. If you choose to use a non-autoinducing media, the concentration of inducer (e.g. isopropyl β-D-1-thiogalactopyranoside (IPTG)) and the timing and length of induction can also significantly influence the yields of folded protein and may need optimization.

Once small-scale screening experiments have concluded positively and you have chosen your expression construct and strain, you may want to scale-up the production and purification of your protein depending on the end use. For an extensive overview of upstream and downstream process development strategies for production of heterologous proteins in *E. coli*, refer to [1,65].

## Summary

- *E. coli* is an excellent host for recombinant protein production in both academia and industry.
- A rational approach is required for successful protein production. Understanding or predicting using bioinformatics tools, the biophysical characteristics of the protein is essential.
- Correct identification of domain boundaries, signal sequences, TM regions, obligate oligomeric complex formation, and PTMs are critical.
- It is equally important to consider genetic and translation factors, such as codon usage, the nature and position of the rbs and differences between prokaryotic and eukaryotic translation rates.
- Other factors such as the strain and media used also impact protein yield, but they cannot compensate for poor planning.

### Competing Interests

The authors declare that there are no competing interests associated with the manuscript.

## Funding

This work was supported the European Union's Horizon 2020 Research and Innovation Programme under Marie Sklodowska-Curie [grant number 642937].

## Author Contribution

L.W.R. conceived the article. All authors contributed to the writing.

## Abbreviations

cDNA, complementary DNA; DNA, deoxyribonucleic acid; IDP, intrinsically disordered protein; LB, Luria–Bertani; MBP, maltose-binding protein; mRNA, messenger RNA; POI, protein of interest; PTM, post-translational modification; rbs, ribosome-binding site; SD, Shine–Dalgarno; SDS/PAGE, sodium dodecyl sulfate/polyacrylamide gel electrophoresis; TM, transmembrane; tRNA, transfer RNA.

# References

1 Tripathi, N.K. and Shrivastava, A. (2019) Recent developments in bioprocessing of recombinant proteins: expression hosts and process development. *Front. Bioeng. Biotechnol.* **7**, 420, https://doi.org/10.3389/fbioe.2019.00420
2 Karyolaimos, A., Ampah-Korsah, H., Zhang, Z. and de Gier, J.-W. (2018) Shaping Escherichia coli for recombinant membrane protein production. *FEMS Microbiol. Lett.* **365**, 152, https://doi.org/10.1093/femsle/fny152
3 Schlegel, S., Hjelm, A., Baumgarten, T., Vikström, D. and de Gier, J.-W.W. (2014) Bacterial-based membrane protein production. *Biochim. Biophys. Acta. Mol. Cell Res.* **1843**, 1739–1749, https://doi.org/10.1016/j.bbamcr.2013.10.023
4 Errey, J.C. and Fiez-Vandal, C. (2020) Production of membrane proteins in industry: the example of GPCRs. *Protein Expr. Purif.* **169**, 105569, https://doi.org/10.1016/j.pep.2020.105569
5 The UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515, https://doi.org/10.1093/nar/gky1049
6 Darlington, P.J., Kirchhof, M.G., Criado, G., Sondhi, J. and Madrenas, J. (2005) Hierarchical regulation of CTLA-4 dimer-based lattice formation and its biological relevance for T cell inactivation. *J. Immunol.* **175**, 996–1004, https://doi.org/10.4049/jimmunol.175.2.996
7 Manta, B., Boyd, D. and Berkmen, M. (2019) Disulfide bond formation in the periplasm of Escherichia coli. *EcoSal Plus* **8**, https://doi.org/10.1128/ecosalplus.ESP-0012-2018
8 Drozdetskiy, A., Cole, C., Procter, J. and Barton, G.J. (2015) JPred4: a protein secondary structure prediction server. *Nucleic Acids Res.* **43**, W389–W394, https://doi.org/10.1093/nar/gkv332
9 Demarest, S.J., Martinez-Yamout, M., Chung, J., Chen, H., Xu, W., Jane Dyson, H. et al. (2002) Mutual synergistic folding in recruitment of cbp/p300 by p160 nuclear receptor coactivators. *Nature* **415**, 549–553, https://doi.org/10.1038/415549a
10 Subramanian, S. and Ross, P.D. (1984) Dye-ligand affinity chromatography: The interaction of cibacron blue f3GA® with proteins and enzyme. *Crit. Rev. Biochem. Mol. Biol.* **16**, 169–205, https://doi.org/10.3109/10409238409102302
11 Kish, W.S., Roach, M.K., Sachi, H., Naik, A.D., Menegatti, S. and Carbonell, R.G. (2018) Purification of human erythropoietin by affinity chromatography using cyclic peptide ligands. *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.* **1085**, 1–12, https://doi.org/10.1016/j.jchromb.2018.03.039
12 Young, C.L., Britton, Z.T. and Robinson, A.S. (2012) Recombinant protein expression and purification: A comprehensive review of affinity tags and microbial applications. *Biotechnol. J.* **7**, 620–634, https://doi.org/10.1002/biot.201100155
13 Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N. and Sternberg, M.J.E. (2015) The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10**, 845–858, https://doi.org/10.1038/nprot.2015.053
14 Geoghegan, K.F., Dixon, H.B.F., Rosner, P.J., Hoth, L.R., Lanzetti, A.J., Borzilleri, K.A. et al. (1999) Spontaneous $\alpha$-N-6-phosphogluconoylation of a "His tag" in Escherichia coli: the cause of extra mass of 258 or 178 Da in fusion proteins. *Anal. Biochem.* **267**, 169–184, https://doi.org/10.1006/abio.1998.2990
15 Castellanos-Serra, L.R., Hardy, E., Ubieta, R., Vispo, N.S., Fernandez, C., Besada, V. et al. (1996) Expression and folding of an interleukin-2-proinsulin fusion protein and its conversion into insulin by a single step enzymatic removal of the C-peptide and the N-terminal fused sequence. *FEBS Lett.* **378**, 171–176, https://doi.org/10.1016/0014-5793(95)01437-3
16 Ludeman, J.P., Pike, R.N., Bromfield, K.M., Duggan, P.J., Cianci, J., Le Bonniec, B. et al. (2003) Determination of the P′1, P′2 and P′3 subsite-specificity of factor Xa. *Int. J. Biochem. Cell Biol.* **35**, 221–225, https://doi.org/10.1016/S1357-2725(02)00128-0
17 Bianchini, E.P., Louvain, V.B., Marque, P.E., Juliano, M.A., Juliano, L. and Le Bonniec, B.F. (2002) Mapping of the catalytic groove preferences of factor Xa reveals an inadequate selectivity for its macromolecule substrates. *J. Biol. Chem.* **277**, 20527–20534, https://doi.org/10.1074/jbc.M201139200
18 Rawlings, N.D., Barrett, A.J., Thomas, P.D., Huang, X., Bateman, A. and Finn, R.D. (2018) The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. *Nucleic Acids Res.* **46**, D624–D632, https://doi.org/10.1093/nar/gkx1134
19 Peng, S., Chu, Z., Lu, J., Li, D., Wang, Y., Yang, S. et al. (2016) Co-expression of chaperones from P. furiosus enhanced the soluble expression of the recombinant hyperthermophilic $\alpha$-amylase in E. coli. *Cell Stress Chaperones* **21**, 477–484, https://doi.org/10.1007/s12192-016-0675-7
20 Amann, T., Schmieder, V., Faustrup Kildegaard, H., Borth, N. and Andersen, M.R. (2019) Genetic engineering approaches to improve posttranslational modification of biopharmaceuticals in different production platforms. *Biotechnol. Bioeng.* **116**, 2778–2796, https://doi.org/10.1002/bit.27101

*Essays in Biochemistry* (2021) **65** 247–260
https://doi.org/10.1042/EBC20200170


21 Alibolandi, M. and Mirzahoseini, H. (2011) Chemical assistance in refolding of bacterial inclusion bodies. *Biochem. Res. Int.* **2011**, 631607, https://doi.org/10.1155/2011/631607

22 Kaur, J.J., Kumar, A. and Kaur, J.J. (2018) Strategies for optimization of heterologous protein expression in E. coli: Roadblocks and reinforcements. *Int. J. Biol. Macromol.* **106**, 803–822, https://doi.org/10.1016/j.ijbiomac.2017.08.080

23 Simmons, L.C. and Yansura, D.G. (1996) Translational level is a critical factor for the secretion of heterologous proteins in Escherichia coli. *Nat. Biotechnol.* **14**, 629–634, https://doi.org/10.1038/nbt0596-629

24 Lobstein, J., Emrich, C.A., Jeans, C., Faulkner, M., Riggs, P. and Berkmen, M. (2012) SHuffle, a novel Escherichia coli protein expression strain capable of correctly folding disulfide bonded proteins in its cytoplasm. *Microb. Cell Fact.* **11**, 1, https://doi.org/10.1186/1475-2859-11-56

25 Bessette, P.H., Åslund, F., Beckwith, J. and Georgiou, G. (1999) Efficient folding of proteins with multiple disulfide bonds in the Escherichia coli cytoplasm. *Proc. Natl. Acad Sci. U.S.A.* **96**, 13703–13708, https://doi.org/10.1073/pnas.96.24.13703

26 Saaranen, M.J. and Ruddock, L.W. (2019) Applications of catalyzed cytoplasmic disulfide bond formation. *Biochem. Soc. Trans.* **47**, 1223–1231, https://doi.org/10.1042/BST20190088

27 Matos, C.F.R.O.R.O., Robinson, C., Alanen, H.I., Prus, P., Uchida, Y., Ruddock, L.W. et al. (2014) Efficient export of prefolded, disulfide-bonded recombinant proteins to the periplasm by the Tat pathway in Escherichia coli CyDisCo strains. *Biotechnol. Prog.* **30**, 281–290, https://doi.org/10.1002/btpr.1858

28 Alanen, H.I., Walker, K.L., Lourdes Velez Suberbie, M., Matos, C.F.R.O., Bönisch, S., Freedman, R.B. et al. (2015) Efficient export of human growth hormone, interferon $\alpha$2b and antibody fragments to the periplasm by the Escherichia coli Tat pathway in the absence of prior disulfide bond formation. *Biochim. Biophys. Acta Mol. Cell Res.* **1853**, 756–763, https://doi.org/10.1016/j.bbamcr.2014.12.027

29 Mueller, P., Gauttam, R., Raab, N., Handrick, R., Wahl, C., Leptihn, S. et al. (2018) High level in vivo mucin-type glycosylation in Escherichia coli. *Microb. Cell Fact.* **17**, 168, https://doi.org/10.1186/s12934-018-1013-9

30 Wingfield, P.T. (2017) N-terminal methionine processing. *Curr. Protoc. Protein Sci.* **88**, 6.14.1–6.14.3, https://doi.org/10.1002/cpps.29

31 Liao, Y.-D., Jeng, J.-C., Wang, C.-F., Wang, S.-C. and Chang, S.-T. (2004) Removal of N-terminal methionine from recombinant proteins by engineered *E. coli* methionine aminopeptidase. *Protein Sci.* **13**, 1802–1810, https://doi.org/10.1110/ps.04679104

32 Tobias, J.W., Shrader, T.E., Rocap, G. and Varshavsky, A. (1991) The N-end rule in bacteria. *Science* **254**, 1374–1377, https://doi.org/10.1126/science.1962196

33 Erbse, A., Schmidt, R., Bornemann, T., Schneider-Mergener, J., Mogk, A., Zahn, R. et al. (2006) ClpS is an essential component of the N-end rule pathway in Escherichia coli. *Nature* **439**, 753–756, https://doi.org/10.1038/nature04412

34 Schuenemann, V.J., Kralik, S.M., Albrecht, R., Spall, S.K., Truscott, K.N., Dougan, D.A. et al. (2009) Structural basis of N-end rule substrate recognition in Escherichia coli by the ClpAP adaptor protein ClpS. *EMBO Rep.* **10**, 508–514, https://doi.org/10.1038/embor.2009.62

35 Celie, P.H.N., Parret, A.H.A. and Perrakis, A. (2016) Recombinant cloning strategies for protein expression. *Curr. Opin. Struct. Biol.* **38**, 145–154, https://doi.org/10.1016/j.sbi.2016.06.010

36 Ou, B., Garcia, C., Wang, Y., Zhang, W. and Zhu, G. (2018) Techniques for chromosomal integration and expression optimization in Escherichia coli. *Biotechnol. Bioeng.* **115**, 2467–2478, https://doi.org/10.1002/bit.26790

37 Stargardt, P., Feuchtenhofer, L., Cserjan-Puschmann, M., Striedner, G. and Mairhofer, J. (2020) Bacteriophage inspired growth-decoupled recombinant protein production in Escherichia coli. *ACS Synth. Biol.* **9**, 1336–1348, https://doi.org/10.1021/acssynbio.0c00028

38 Chen, H., Bjerknes, M., Kumar, R. and Jay, E. (1994) Determination of the optimal aligned spacing between the shine - dalgarno sequence and the translation initiation codon of escherichia coli m RNAs. *Nucleic Acids Res.* **22**, 4953–4957, https://doi.org/10.1093/nar/22.23.4953

39 Shepard, H.M., Yelverton, E. and Goeddel, D.V. (1982) Increased synthesis in E. coli of fibroblast and leukocyte interferons through alterations in ribosome binding sites. *DNA* **1**, 125–131, https://doi.org/10.1089/dna.1.1982.1.125

40 Shine, J. and Dalgarno, L. (1974) The 3′ terminal sequence of Escherichia coli 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc. Natl. Acad. Sci. U.S.A.* **71**, 1342–1346, https://doi.org/10.1073/pnas.71.4.1342

41 Ikemura, T. (1981) Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes. *J. Mol. Biol.* **146**, 1–21, https://doi.org/10.1016/0022-2836(81)90363-6

42 Boël, G., Letso, R., Neely, H., Price, W.N., Wong, K.H., Su, M. et al. (2016) Codon influence on protein expression in E. coli correlates with mRNA levels. *Nature* **529**, 358–363, https://doi.org/10.1038/nature16509

43 Fuhrmann, M., Hausherr, A., Ferbitz, L., Schödl, T., Heitzer, M. and Hegemann, P. (2004) Monitoring dynamic expression of nuclear genes in Chlamydomonas reinhardtii by using a synthetic luciferase reporter gene. *Plant Mol. Biol.* **55**, 869–881, https://doi.org/10.1007/s11103-005-2150-1

44 Kleber-Janke, T. and Becker, W.M. (2000) Use of modified BL21(DE3) Escherichia coli cells for high-level expression of recombinant peanut allergens affected by poor codon usage. *Protein Expr. Purif.* **19**, 419–424, https://doi.org/10.1006/prep.2000.1265

45 Lipinszki, Z., Vernyik, V., Farago, N., Sari, T., Puskas, L.G., Blattner, F.R. et al. (2018) Enhancing the translational capacity of E. coli by resolving the codon bias. *ACS Synth. Biol.* **7**, 2656–2664, https://doi.org/10.1021/acssynbio.8b00332

46 Novy, R., Drott, D., Yaeger, K. and Mierendorf, R. (2001) Overcoming the codon bias of *E. coli* for enhanced protein expression. *inNovations* **12**, 1–3

47 Komar, A.A. (2016) The Yin and Yang of codon usage. *Hum. Mol. Genet.* **25**, R77–R85, https://doi.org/10.1093/hmg/ddw207

48 Chemla, Y., Peeri, M., Heltberg, M.L., Eichler, J., Jensen, M.H., Tuller, T. et al. (2020) A possible universal role for mRNA secondary structure in bacterial translation revealed using a synthetic operon. *Nat. Commun.* **11**, 1–11, https://doi.org/10.1038/s41467-020-18577-4

49 Lenz, G., Doron-Faigenboim, A., Ron, E.Z., Tuller, T. and Gophna, U. (2011) Sequence features of E. coli mRNAs affect their degradation. *PLoS ONE* **6**, e28544, https://doi.org/10.1371/journal.pone.0028544

50 Dennis, P.P. and Bremer, H. (2008) Modulation of chemical composition and other parameters of the cell at different exponential growth rates. *EcoSal Plus* **3**, https://doi.org/10.1128/ecosal.5.2.3

51 Riba, A., Di Nanni, N., Mittal, N., Arhné, E., Schmidt, A. and Zavolan, M. (2019) Protein synthesis rates and ribosome occupancies reveal determinants of translation elongation rates. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 15023LP–15032LP, https://doi.org/10.1073/pnas.1817299116
52 Siller, E., DeZwaan, D.C., Anderson, J.F., Freeman, B.C. and Barral, J.M. (2010) Slowing bacterial translation speed enhances eukaryotic protein folding efficiency. *J. Mol. Biol.* **396**, 1310–1318, https://doi.org/10.1016/j.jmb.2009.12.042
53 Angov, E., Hillier, C.J., Kincaid, R.L. and Lyon, J.A. (2008) Heterologous protein expression is enhanced by harmonizing the codon usage frequencies of the target gene with those of the expression host. *PLoS ONE* **3**, e2189, https://doi.org/10.1371/journal.pone.0002189
54 Zhang, G. and Ignatova, Z. (2011) Folding at the birth of the nascent chain: coordinating translation with co-translational folding. *Curr. Opin. Struct. Biol.* **21**, 25–31, https://doi.org/10.1016/j.sbi.2010.10.008
55 Hui, A. and De Boer, H.A. (1987) Specialized ribosome system: preferential translation of a single mRNA species by a subpopulation of mutated ribosomes in Escherichia coli. *Proc. Natl. Acad. Sci. U.S.A.* **84**, 4762–4766, https://doi.org/10.1073/pnas.84.14.4762
56 Dixon, N., Robinson, C.J., Geerlings, T., Duncan, J.N., Drummond, S.P. and Micklefield, J. (2012) Orthogonal riboswitches for tuneable coexpression in bacteria. *Angew Chemie. Int. Ed.* **51**, 3620–3624, https://doi.org/10.1002/anie.201109106
57 Morra, R., Shankar, J., Robinson, C.J., Halliwell, S., Butler, L., Upton, M. et al. (2016) Dual transcriptional- Translational cascade permits cellular level tuneable expression control. *Nucleic Acids Res.* **44**, 21, https://doi.org/10.1093/nar/gkv912
58 Orelle, C., Carlson, E.D., Szal, T., Florin, T., Jewett, M.C. and Mankin, A.S. (2015) Protein synthesis by ribosomes with tethered subunits. *Nature* **524**, 119–124, https://doi.org/10.1038/nature14862
59 Horga, L.G., Halliwell, S., Castiñeiras, T.S., Wyre, C., Matos, C.F.R.O., Yovcheva, D.S. et al. (2018) Tuning recombinant protein expression to match secretion capacity. *Microb. Cell Fact.* **17**, 199, https://doi.org/10.1186/s12934-018-1047-z
60 Carlson, E.D., d'Aquino, A.E., Kim, D.S., Fulk, E.M., Hoang, K., Szal, T. et al. (2019) Engineered ribosomes with tethered subunits for expanding biological function. *Nat. Commun.* **10**, 1–13, https://doi.org/10.1038/s41467-019-11427-y
61 Lukjancenko, O., Wassenaar, T.M. and Ussery, D.W. (2010) Comparison of 61 sequenced Escherichia coli genomes. *Microb. Ecol.* **60**, 708–720, 2010/07/11, https://doi.org/10.1007/s00248-010-9717-3
62 Studier, F.W. (2005) Protein production by auto-induction in high density shaking cultures. *Protein Expr. Purif.* **41**, 207–234, https://doi.org/10.1016/j.pep.2005.01.016
63 Ukkonen, K., Neubauer, A., Pereira, V.J. and Vasala, A. (2017) High yield of recombinant protein in shaken *E. coli* cultures with enzymatic glucose release medium EnPresso B. *Methods Mol. Biol.* **1586**, 127–137, https://doi.org/10.1007/978-1-4939-6887-9˙8
64 Rosano, G.L. and Ceccarelli, E.A. (2014) Recombinant protein expression in Escherichia coli: advances and challenges. *Front. Microbiol.* **5**, 172, https://doi.org/10.3389/fmicb.2014.00172
65 Tripathi, N.K. (2016) Production and purification of recombinant proteins from *Escherichia coli*. *Chem. Biol. Eng. Rev.* **3**, 116–133, https://doi.org/10.1002/cben.201600002
66 Artimo, P., Jonnalagedda, M., Arnold, K., Baratin, D., Csardi, G., de Castro, E. et al. (2012) ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Res.* **40**, W597–W603, https://doi.org/10.1093/nar/gks400
67 Terpe, K. (2006) Overview of bacterial expression systems for heterologous protein production: from molecular and biochemical fundamentals to commercial systems. *Appl. Microbiol. Biotechnol.* **72**, 211–222, https://doi.org/10.1007/s00253-006-0465-8
68 Marschall, L., Sagmeister, P. and Herwig, C. (2017) Tunable recombinant protein expression in E. coli: promoter systems and genetic constraints. *Appl. Microbiol. Biotechnol.* **101**, 501–512, https://doi.org/10.1007/s00253-016-8045-z
69 Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. et al. (2013) GenBank. *Nucleic Acids Res.* **41**, D36–D42, https://doi.org/10.1093/nar/gks1195
70 Almagro Armenteros, J.J., Tsirigos, K.D., Sønderby, C.K., Petersen, T.N., Winther, O., Brunak, S. et al. (2019) SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* **37**, 420–423, https://doi.org/10.1038/s41587-019-0036-z
71 Freudl, R. (2018) Signal peptides for recombinant protein secretion in bacterial expression systems. *Microb. Cell Fact.* **17**, https://doi.org/10.1186/s12934-018-0901-3
72 Tsirigos, K.D., Peters, C., Shu, N., Käll, L. and Elofsson, A. (2015) The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. *Nucleic Acids Res.* **43**, W401–W407, https://doi.org/10.1093/nar/gkv485
73 Mazola, Y., Chinea, G. and Musacchio, A. (2011) Integrating bioinformatics tools to handle glycosylation. *PLoS Comput. Biol.* **7**, e1002285, https://doi.org/10.1371/journal.pcbi.1002285
74 Monigatti, F., Gasteiger, E., Bairoch, A. and Jung, E. (2002) The Sulfinator: predicting tyrosine sulfation sites in protein sequences. *Bioinformatics* **18**, 769–770, https://doi.org/10.1093/bioinformatics/18.5.769
75 Blom, N., Gammeltoft, S. and Brunak, S. (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol.* **294**, 1351–1362, https://doi.org/10.1006/jmbi.1999.3310
76 Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R.D. and Bairoch, A. (2003) ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.* **31**, 3784–3788, https://doi.org/10.1093/nar/gkg563
77 Bolivar, F., Rodriguez, R.L., Betlach, M.C. and Boyer, H.W. (1977) Construction and characterization of new cloning vehicles. I. Ampicillin-resistant derivatives of the plasmid pMB9. *Gene* **2**, 75–93, https://doi.org/10.1016/0378-1119(77)90074-9
78 Vieira, J. and Messing, J. (1982) The pUC plasmids, an M13mp7-derived system for insertion mutagenesis and sequencing with synthetic universal primers. *Gene* **19**, 259–268, https://doi.org/10.1016/0378-1119(82)90015-4
79 Yanisch-Perron, C., Vieira, J. and Messing, J. (1985) Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mpl8 and pUC19 vectors. *Gene* **33**, 103–119, https://doi.org/10.1016/0378-1119(85)90120-9
80 Hershfield, V., Boyer, H.W., Yanofsky, C., Lovett, M.A. and Helinski, D.R. (1974) Plasmid ColEl as a molecular vehicle for cloning and amplification of DNA. *Proc. Natl. Acad. Sci. U.S.A.* **71**, 3455–3459, https://doi.org/10.1073/pnas.71.9.3455

*Essays in Biochemistry* (2021) **65** 247–260
https://doi.org/10.1042/EBC20200170

81 Eun, H.-M. (1996) Marker/reporter enzymes. *Enzymology Primer for Recombinant DNA Technology*, Academic Press, San Diego
82 Chang, A.C. and Cohen, S.N. (1978) Construction and characterization of amplifiable multicopy DNA cloning vehicles derived from the P15A cryptic miniplasmid. *J. Bacteriol.* **134**, 1141–1156, https://doi.org/10.1128/JB.134.3.1141-1156.1978
83 Shafferman, A. and Helinski, D.R. (1983) Structural properties of the beta origin of replication of plasmid R6K. *J. Biol. Chem.* **258**, 4083–4090, https://doi.org/10.1016/S0021-9258(18)32587-0
84 Cohen, S.N. and Chang, A.C.Y. (1977) Revised interpretation of the origin of the pSC101 plasmid. *J. Bacteriol.* **132**, https://doi.org/10.1128/JB.132.2.734-737.1977
85 Hasunuma, K. and Sekiguchi, M. (1977) Replication of plasmid pSC101 in Escherichia coli K12: requirement for dnaA function. *Mol. Gen Genet.* **154**, 225–230, https://doi.org/10.1007/BF00571277
86 Sutcliffe, J.G. (1978) Nucleotide sequence of the ampicillin resistance gene of Escherichia coli plasmid pBR322. *Proc. Natl. Acad. Sci. U.S.A.* **75**, 3737–3741, https://doi.org/10.1073/pnas.75.8.3737
87 Schwarz, S., Kehrenberg, C., Doublet, B.B. and Cloeckaert, A. (2004) Molecular basis of bacterial resistance to chloramphenicol and florfenicol. *FEMS Microbiol. Rev.* **28**, 519–542, https://doi.org/10.1016/j.femsre.2004.04.001
88 Jelenić, S. (2003) Controversy associated with the common component of most transgenic plants - Kanamycin resistance marker gene. *Food Technol. Biotechnol.* **41**, 183–190
89 Møller, T.S.B., Overgaard, M., Nielsen, S.S., Bortolaia, V., Sommer, M.O.A., Guardabassi, L. et al. (2016) Relation between tetR and tetA expression in tetracycline resistant Escherichia coli. *BMC Microbiol.* **16**, 39, https://doi.org/10.1186/s12866-016-0649-z
90 Ramirez, M.S. and Tolmasky, M.E. (2010) Aminoglycoside modifying enzymes. *Drug Resist. Updat.* **13**, 151–171, https://doi.org/10.1016/j.drup.2010.08.003
91 Ali, S.A. and Chew, Y.W. (2015) FabV/triclosan is an antibiotic-free and cost-effective selection system for efficient maintenance of high and medium-copy number plasmids in Escherichia coli. *PLoS ONE* **10**, e0129547, https://doi.org/10.1371/journal.pone.0129547
92 Fiedler, M. and Skerra, A. (2001) proBA complementation of an auxotrophic E. coli strain improves plasmid stability and expression yield during fermenter production of a recombinant antibody fragment. *Gene* **274**, 111–118, https://doi.org/10.1016/S0378-1119(01)00629-1
93 Velur Selvamani, R.S., Telaar, M., Friehs, K. and Flaschel, E. (2014) Antibiotic-free segregational plasmid stabilization in Escherichia coli owing to the knockout of triosephosphate isomerase (tpiA). *Microb. Cell Fact.* **13**, 58, https://doi.org/10.1186/1475-2859-13-58
94 Vidal, L., Pinsach, J., Striedner, G., Caminal, G. and Ferrer, P. (2008) Development of an antibiotic-free plasmid selection system based on glycine auxotrophy for recombinant protein overproduction in Escherichia coli. *J. Biotechnol.* **134**, 127–136, https://doi.org/10.1016/j.jbiotec.2008.01.011
95 Dong, W.R., Xiang, L.X. and Shao, J.Z. (2010) Novel antibiotic-free plasmid selection system based on complementation of host auxotrophy in the NAD de novo synthesis pathway. *Appl. Environ. Microbiol.* **76**, 2295–2303, https://doi.org/10.1128/AEM.02462-09
96 Cranenburgh, R.M., Lewis, K.S. and Hanak, J.A.J. (2004) Effect of plasmid copy number and lac operator sequence on antibiotic-free plasmid selection by operator-repressor titration in Escherichia coli. *J. Mol. Microbiol. Biotechnol.* **7**, 197–203, https://doi.org/10.1159/000079828
97 Ohashi-Kunihiro, S., Hagiwara, H., Yohda, M., Masaki, H. and Machida, M. (2006) Construction of a positive selection marker by a lethal gene with the amber stop codon(s) regulator. *Biosci. Biotechnol. Biochem.* **70**, 119–125, https://doi.org/10.1271/bbb.70.119
98 Rosano, G.L., Morales, E.S. and Ceccarelli, E.A. (2019) New tools for recombinant protein production in Escherichia coli: a 5-year update. *Protein Sci.* **28**, 1412–1422, https://doi.org/10.1002/pro.3668
99 Dumon-Seignovert, L., Cariot, G. and Vuillard, L. (2004) The toxicity of recombinant proteins in Escherichia coli: a comparison of overexpression in BL21(DE3), C41(DE3), and C43(DE3). *Protein Expr. Purif.* **37**, 203–206, https://doi.org/10.1016/j.pep.2004.04.025
100 Vijayendran, C., Polen, T., Wendisch, V.F., Friehs, K., Niehaus, K. and Flaschel, E. (2007) The plasticity of global proteome and genome expression analyzed in closely related W3110 and MG1655 strains of a well-studied model organism, Escherichia coli-K12. *J. Biotechnol.* **128**, 747–761, https://doi.org/10.1016/j.jbiotec.2006.12.026
101 Marisch, K., Bayer, K., Scharl, T., Mairhofer, J., Krempl, P.M., Hummel, K. et al. (2013) A comparative analysis of industrial Escherichia coli K-12 and B strains in high-glucose batch cultivations on process-, transcriptome- and proteome level. *PLoS ONE* **8**, e70516, https://doi.org/10.1371/journal.pone.0070516