So you want to express your protein in *E. 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, 90220 Oulu, Finland

* Communicating author. Email: lloyd.ruddock@oulu.fi

† Authors contributed equally

Abstract

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 and designing expression constructs, as well as insights into recently developed technologies that have contributed to this ever-growing field.

Keywords: bacterial expression, *Escherichia coli*, genetic engineering, recombinant proteins, synthetic biology

Introduction

Ever since the FDA approved the first recombinant protein for therapeutic use in 1982, *E. 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.

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 bioinformatic analysis (Table 1).

While bioinformatic approaches are powerful, they are only predictions and so gathering a consensus from multiple independent bioinformatic approaches or looking for validation through experimental means (e.g., from published literature) is always worthwhile. For example, human CTLA-4 is an obligate dimer and requires N-glycosylation of Asn78 and Asn110 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 NCBD) [9] and so, co-expression of a partner may help considerably in obtaining the protein in a stable and soluble form.

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-terminus 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 to 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 protein of interest 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 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 wild-type *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].

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* [20,21]. 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 [22]. Thirdly, an engineered strain could be used that removes disulfide bond reducing pathways from the cytoplasm [23,24] , or adds oxidative folding catalysts, reviewed in [25]. This can be combined with the TAT-secretion system for exporting folded proteins to the periplasm e.g. [26,27]. Similar synthetic biology approaches also allow other PTMs to be made in the cytoplasm, for example, mucin-type O-glycosylation in *E. coli*. [28].

Finally, it should be remembered that the cytoplasm of *E. coli* contains methionine aminopeptidase, which can remove the initiating methionine [29], 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. [30]. 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 [31], but this depends on the context of the N-terminal and subsequent amino acids [32,33].

After all these considerations, if no purified protein is obtained, a simple troubleshooting SDS-PAGE analysis may quickly help elucidate the possible issues (Figure 1).

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 DNA and mRNA, and the mRNA produced is in far excess of the limitations of the translation apparatus (Figure 2).

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 2) 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* RNA polymerase inhibitor peptide called Gp2 [34]. 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 cDNA library obtained by 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.

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 sequence or ribosome binding site (rbs; Figure 3). The rbs are usually 5-13 base pairs [35] upstream of the initiating AUG (optimal distance 5-6 base pairs [36]); and are complementary to the 3' end of the 16S ribosomal RNA. In *E. coli*, this sequence is AGGAGGU [37]. The requirement for a distinct rbs has two consequences for eukaryotic protein expression in *E. coli*. Firstly, a rbs must be present before the initiating AUG. This may be present in the plasmid outside the multi-cloning 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 amongst 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 tRNA levels [38]. mRNA which contains multiple rare codons can exhibit translation stalling and mRNA degradation, reviewed in [39]. Codon usage issues can be examined by bioinformatic approaches e.g., Graphical Codon Usage Analyzer [40]. One method to prevent this problem was the overexpression of rare tRNAs e.g., [41,42] such as from pLysSRARE [43]. For more detailed insights into codon usage, refer to [44]. 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 [45,46]. As prices have rapidly dropped a synthetic gene can cost less than the labour 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 circa 50 nucleotides/s and translation rates circa 16 amino acids/s [47]. In contrast translation rates in eukaryotes are slower, with a rate of circa 3 amino acids/s [48]. 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 multi-domain proteins, this can be a serious issue (Figure 4). This can be mitigated by modulation of translation rate [49], codon usage harmonization [50], or the use of rarer codons just after domain boundaries to cause ribosome stalling [51] (Figure 4).

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 Anna Hui and Herman A.D. Boer in 1987 [52]. Alternative ribosome systems such as the orthogonal riboswitch system [53], the *RiboTite* system [54], and the Ribo-T system [55] have been reported since. The riboswitch system allows tunable coexpression 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 [56]. 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 [55]. This orthogonal ribosome-mRNA system is capable of supporting bacterial growth even in the absence of wild-type ribosomes and its improved tethered version has been reported recently [57].

One other 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 a rbs before the initiating AUG of each (Figure 3). 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.

Strain and media

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 circa 20% of the genome common to all strains [58]. It can be broadly split into 4 sub-groupings, K-12 strains, B-strains, and the C and W strain based on their initial isolation [58]. 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 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 autoinduction media e.g. [59] both facilitates the screening of multiple POI and allows culture densities typically 10x higher than LB.

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 [60] so as not to exceed the capacity of the folding machinery.

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, transmembrane regions, obligate oligomeric complex formation and post-translational modifications are critical.
- It is equally important to consider genetic and translation factors, such as codon usage, the nature and position of the ribosome binding site 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

The authors receive funding from the European Union's Horizon 2020 research and innovation programme under Marie Sklodowska-Curie grant agreement No 642937.

Author contributions

LWR conceived the article, all authors contributed to the writing.

Abbreviations

cDNA: complementary DNA

CTLA-4: Cytotoxic T-Lymphocyte antigen 4

DNA: Deoxyribonucleic acid

FDA: Food and Drug Administration

LB: Luria-Bertani

mRNA: messenger RNA

NCBD: Nuclear co-activator binding domain

pI: Isoelectric point

RNA: Ribonucleic acid

RT-PCR: Real Time-Polymerase Chain Reaction

SDS-PAGE: Sodium dodecyl sulfate-Polyacrylamide Gel Electrophoresis

tRNA: transfer RNA

Figure Legends:

Figure 1. 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 protein of interest 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.

Figure 2. 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.

Figure 3. Schematic representation of initiation of translation in prokaryotes and eukaryotes.
A. The process of translation is carried out by a ribosome comprised of 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 Shine-Dalgarno (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.

Figure 4. 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.

References:

1. Tripathi NK, 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, 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, de Gier J-WW. (2014). Bacterial-based membrane protein production. *Biochim Biophys Acta - Mol Cell Res*. **1843**:1739–49. https://doi.org/https://doi.org/10.1016/j.bbamcr.2013.10.023

4. Errey JC, 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**:D506–15. https://doi.org/10.1093/nar/gky1049

6. Darlington PJ, Kirchhof MG, Criado G, Sondhi J, 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, 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, Barton GJ. (2015). JPred4: A protein secondary structure prediction server. *Nucleic Acids Res*. **43**:W389–94. https://doi.org/10.1093/nar/gkv332

9. Demarest SJ, 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**:549–53. https://doi.org/10.1038/415549a

10. Subramanian S, Ross PD. (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 WS, Roach MK, Sachi H, Naik AD, Menegatti S, Carbonell RG. (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 CL, Britton ZT, Robinson AS. (2012). Recombinant protein expression and purification: A comprehensive review of affinity tags and microbial applications. *Biotechnol J*. **7**:620–34. https://doi.org/10.1002/biot.201100155

13. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc*. **10**:845–58. https://doi.org/10.1038/nprot.2015.053

14. Geoghegan KF, Dixon HBF, Rosner PJ, Hoth LR, Lanzetti AJ, Borzilleri KA, et al. (1999). Spontaneous α-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**:169–84. https://doi.org/10.1006/abio.1998.2990

15. Castellanos-Serra LR, Hardy E, Ubieta R, Vispo NS, 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**:171–6. https://doi.org/10.1016/0014-5793(95)01437-3

16. Ludeman JP, Pike RN, Bromfield KM, Duggan PJ, 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**:221–5. https://doi.org/10.1016/S1357-2725(02)00128-0

17. Bianchini EP, Louvain VB, Marque PE, Juliano MA, Juliano L, Le Bonniec BF. (2002). Mapping of the catalytic groove preferences of factor Xa reveals an inadequate selectivity for its macromolecule substrates. *J Biol Chem*. **277**:20527–34. https://doi.org/10.1074/jbc.M201139200

18. Rawlings ND, Barrett AJ, Thomas PD, Huang X, Bateman A, Finn RD. (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**:D624–32. 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 α-amylase in E. coli. *Cell Stress Chaperones*. **21**:477–84. https://doi.org/10.1007/s12192-016-0675-7

20. Alibolandi M, Mirzahoseini H. (2011). Chemical Assistance in Refolding of Bacterial Inclusion Bodies. *Biochem Res Int*. **2011**:631607. https://doi.org/10.1155/2011/631607

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

22. Simmons LC, Yansura DG. (1996). Translational level is a critical factor for the secretion of heterologous proteins in Escherichia coli. *Nat Biotechnol*. **14**:629–34. https://doi.org/10.1038/nbt0596-629

23. Lobstein J, Emrich CA, Jeans C, Faulkner M, Riggs P, 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

24. Bessette PH, Åslund F, Beckwith J, Georgiou G. (1999). Efficient folding of proteins with multiple disulfide bonds in the Escherichia coli cytoplasm. *Proc Natl Acad Sci U S A*. **96**:13703–8. https://doi.org/10.1073/pnas.96.24.13703

25. Saaranen MJ, Ruddock LW. (2019). Applications of catalyzed cytoplasmic disulfide bond formation. *Biochem Soc Trans*. **47**:1223–31. https://doi.org/10.1042/BST20190088

26. Matos CFRORO, Robinson C, Alanen HI, Prus P, Uchida Y, Ruddock LW, 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**:281–90. https://doi.org/10.1002/btpr.1858

27. Alanen HI, Walker KL, Lourdes Velez Suberbie M, Matos CFRO, Bönisch S, Freedman RB, et al. (2015). Efficient export of human growth hormone, interferon α2b 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**:756–63. https://doi.org/10.1016/j.bbamcr.2014.12.027

28. 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

29. Wingfield PT. (2017). N-Terminal Methionine Processing. *Curr Protoc Protein Sci*. **88**:6.14.1-6.14.3. https://doi.org/10.1002/cpps.29

30. Liao Y-D, Jeng J-C, Wang C-F, Wang S-C, Chang S-T. (2004). Removal of N-terminal methionine from recombinant proteins by engineered *E. coli* methionine aminopeptidase. *Protein Sci*. **13**:1802–10. https://doi.org/10.1110/ps.04679104

31. Tobias JW, Shrader TE, Rocap G, Varshavsky A. (1991). The N-end rule in bacteria. *Science (80-)*. **254**:1374–7. https://doi.org/10.1126/science.1962196

32. 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**:753–6. https://doi.org/10.1038/nature04412

33. Schuenemann VJ, Kralik SM, Albrecht R, Spall SK, Truscott KN, Dougan DA, et al. (2009). Structural basis of N-end rule substrate recognition in Escherichia coli by the ClpAP adaptor protein ClpS. *EMBO Rep*. **10**:508–14. https://doi.org/10.1038/embor.2009.62

34. Stargardt P, Feuchtenhofer L, Cserjan-Puschmann M, Striedner G, Mairhofer J. (2020). Bacteriophage Inspired Growth-Decoupled Recombinant Protein Production in Escherichia coli. *ACS Synth Biol*. **9**:1336–48. https://doi.org/10.1021/acssynbio.0c00028

35. Chen H, Bjerknes M, Kumar R, 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**:4953–7. https://doi.org/10.1093/nar/22.23.4953

36. Shepard HM, Yelverton E, Goeddel D V. (1982). Increased Synthesis in E. coli of Fibroblast and Leukocyte Interferons Through Alterations in Ribosome Binding Sites. *DNA*. **1**:125–31. https://doi.org/10.1089/dna.1.1982.1.125

37. Shine J, 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**:1342–6. https://doi.org/10.1073/pnas.71.4.1342

38. 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

39. Boël G, Letso R, Neely H, Price WN, Wong KH, Su M, et al. (2016). Codon influence on protein expression in E. coli correlates with mRNA levels. *Nature*. **529**:358–63. https://doi.org/10.1038/nature16509

40. Fuhrmann M, Hausherr A, Ferbitz L, Schödl T, Heitzer M, Hegemann P. (2004). Monitoring dynamic expression of nuclear genes in Chlamydomonas reinhardtii by using a synthetic luciferase reporter gene. *Plant Mol Biol*. **55**:869–81. https://doi.org/10.1007/s11103-005-2150-1

41. Kleber-Janke T, Becker WM. (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**:419–24. https://doi.org/10.1006/prep.2000.1265

42. Lipinszki Z, Vernyik V, Farago N, Sari T, Puskas LG, Blattner FR, et al. (2018). Enhancing the Translational Capacity of E. coli by Resolving the Codon Bias. https://doi.org/10.1021/acssynbio.8b00332

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

44. Komar AA. (2016). The Yin and Yang of codon usage. *Hum Mol Genet*. **25**:R77–85. https://doi.org/10.1093/hmg/ddw207

45. Chemla Y, Peeri M, Heltberg ML, Eichler J, Jensen MH, 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

46. Lenz G, Doron-Faigenboim A, Ron EZ, Tuller T, Gophna U. (2011). Sequence Features of E. coli mRNAs Affect Their Degradation. *PLoS One*. **6**:e28544. https://doi.org/10.1371/journal.pone.0028544

47. Dennis PP, 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

48. Riba A, Di Nanni N, Mittal N, Arhné E, Schmidt A, Zavolan M. (2019). Protein synthesis rates and ribosome occupancies reveal determinants of translation elongation rates. *Proc Natl Acad Sci*. **116**:15023 LP – 15032. https://doi.org/10.1073/pnas.1817299116

49. Siller E, DeZwaan DC, Anderson JF, Freeman BC, Barral JM. (2010). Slowing Bacterial Translation Speed Enhances Eukaryotic Protein Folding Efficiency. *J Mol Biol*. **396**:1310–8. https://doi.org/10.1016/j.jmb.2009.12.042

50. Angov E, Hillier CJ, Kincaid RL, Lyon JA. (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

51. Zhang G, Ignatova Z. (2011.). Folding at the birth of the nascent chain: Coordinating translation with co-translational folding. Curr. Opin. Struct. Biol. p. 25–31. https://doi.org/10.1016/j.sbi.2010.10.008

52. Hui A, De Boer HA. (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**:4762–6. https://doi.org/10.1073/pnas.84.14.4762

53. Dixon N, Robinson CJ, Geerlings T, Duncan JN, Drummond SP, Micklefield J. (2012). Orthogonal Riboswitches for Tuneable Coexpression in Bacteria. *Angew Chemie Int Ed*. **51**:3620–4. https://doi.org/10.1002/anie.201109106

54. Morra R, Shankar J, Robinson CJ, 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

55. Orelle C, Carlson ED, Szal T, Florin T, Jewett MC, Mankin AS. (2015). Protein synthesis by ribosomes with tethered subunits. *Nature*. **524**:119–24. https://doi.org/10.1038/nature14862

56. Horga LG, Halliwell S, Castiñeiras TS, Wyre C, Matos CFRO, Yovcheva DS, 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

57. Carlson ED, d'Aquino AE, Kim DS, Fulk EM, 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

58. Lukjancenko O, Wassenaar TM, Ussery DW. (2010). Comparison of 61 sequenced Escherichia coli genomes. *Microb Ecol*. 2010/07/11. **60**:708–20. https://doi.org/10.1007/s00248-010-9717-3

59. Studier FW. (2005). Protein production by auto-induction in high density shaking cultures. *Protein Expr Purif*. **41**:207–34. https://doi.org/10.1016/j.pep.2005.01.016

60. Rosano GL, Ceccarelli EA. (2014). Recombinant protein expression in Escherichia coli: advances and challenges. *Front Microbiol*. **5**:172. https://doi.org/10.3389/fmicb.2014.00172

61. 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**:W597-603. https://doi.org/10.1093/nar/gks400

62. Terpe K. (2006). Overview of bacterial expression systems for heterologous protein production: From molecular and biochemical fundamentals to commercial systems. *Appl Microbiol Biotechnol*. **72**:211–22. https://doi.org/10.1007/s00253-006-0465-8

63. Marschall L, Sagmeister P, Herwig C. (2017). Tunable recombinant protein expression in E. coli: promoter systems and genetic constraints. *Appl Microbiol Biotechnol*. **101**:501–12.

https://doi.org/10.1007/s00253-016-8045-z

64. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, et al. (2013). GenBank. *Nucleic Acids Res*. **41**:D36-42. https://doi.org/10.1093/nar/gks1195

65. Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, et al. (2019). SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat Biotechnol*. **37**:420–3. https://doi.org/10.1038/s41587-019-0036-z

66. Freudl R. (2018.). Signal peptides for recombinant protein secretion in bacterial expression systems. Microb. Cell Fact. https://doi.org/10.1186/s12934-018-0901-3

67. Tsirigos KD, Peters C, Shu N, Käll L, Elofsson A. (2015). The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. *Nucleic Acids Res*. **43**:W401–7. https://doi.org/10.1093/nar/gkv485

68. Mazola Y, Chinea G, Musacchio A. (2011). Integrating Bioinformatics Tools to Handle Glycosylation. *PLOS Comput Biol*. **7**:e1002285. https://doi.org/10.1371/journal.pcbi.1002285

69. Monigatti F, Gasteiger E, Bairoch A, Jung E. (2002). The Sulfinator: Predicting tyrosine sulfation sites in protein sequences. *Bioinformatics*. **18**:769–70. https://doi.org/10.1093/bioinformatics/18.5.769

70. Blom N, Gammeltoft S, Brunak S. (1999). Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J Mol Biol*. **294**:1351–62. https://doi.org/10.1006/jmbi.1999.3310

71. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. (2003). ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res*. **31**:3784–8. https://doi.org/10.1093/nar/gkg563

72. Bolivar F, Rodriguez RL, Betlach MC, Boyer HW. (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

73. Vieira J, Messing J. (1982). The pUC plasmids, an M13mp7-derived system for insertion mutagenesis and sequencing with synthetic universal primers. *Gene*. **19**:259–68. https://doi.org/10.1016/0378-1119(82)90015-4

74. Yanisch-Perron C, Vieira J, Messing J. (1985). Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mpl8 and pUC19 vectors. *Gene*. **33**:103–19. https://doi.org/10.1016/0378-1119(85)90120-9

75. Hershfield V, Boyer HW, Yanofsky C, Lovett MA, Helinski DR. (1974). Plasmid ColEl as a molecular vehicle for cloning and amplification of DNA. *Proc Natl Acad Sci U S A*. **71**:3455–9. https://doi.org/10.1073/pnas.71.9.3455

76. Eun H-M. (1996.). Marker/Reporter Enzymes. *Enzymol Prim Recomb DNA Technol*. p. 567–645. https://doi.org/10.1016/b978-012243740-3/50011-9

77. Chang AC, Cohen SN. (1978). Construction and characterization of amplifiable multicopy DNA cloning vehicles derived from the P15A cryptic miniplasmid. *J Bacteriol*. **134**.

78. Shafferman A, Helinski DR. (1983). Structural properties of the beta origin of replication of plasmid R6K. *J Biol Chem*. **258**:4083–90. https://doi.org/https://doi.org/10.1016/S0021-9258(18)32587-0

79. Cohen SN, Chang ACY. (1977). Revised Interpretation of the Origin of the pSC101 Plasmid. *J Bacteriol*. **132**.

80. Hasunuma K, Sekiguchi M. (1977). Replication of plasmid pSC101 in Escherichia coli K12:

Requirement for dnaA function. *MGG Mol Gen Genet*. **154**:225–30. https://doi.org/10.1007/BF00571277

81. Sutcliffe JG. (1978). Nucleotide sequence of the ampicillin resistance gene of Escherichia coli plasmid pBR322. *Proc Natl Acad Sci U S A*. **75**:3737–41. https://doi.org/10.1073/pnas.75.8.3737

82. Schwarz S, Kehrenberg C, Doublet BB, Cloeckaert A. (2004). Molecular basis of bacterial resistance to chloramphenicol and florfenicol. *FEMS Microbiol Rev*. **28**:519–42. https://doi.org/https://doi.org/10.1016/j.femsre.2004.04.001

83. Jelenić S. (2003). Controversy associated with the common component of most transgenic plants - Kanamycin resistance marker gene. *Food Technol Biotechnol*. **41**:183–90.

84. Møller TSB, Overgaard M, Nielsen SS, Bortolaia V, Sommer MOA, 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

85. Ramirez MS, Tolmasky ME. (2010). Aminoglycoside modifying enzymes. *Drug Resist Updat*. **13**:151–71. https://doi.org/https://doi.org/10.1016/j.drup.2010.08.003

86. Ali SA, Chew YW. (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

87. Fiedler M, 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**:111–8. https://doi.org/10.1016/S0378-1119(01)00629-1

88. Velur Selvamani RS, Telaar M, Friehs K, 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

89. Vidal L, Pinsach J, Striedner G, Caminal G, 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**:127–36. https://doi.org/10.1016/j.jbiotec.2008.01.011

90. Dong WR, Xiang LX, Shao JZ. (2010). Novel antibiotic-free plasmid selection system based on complementation of host auxotrophy in the NAD de novo synthesis pathway. *Appl Environ Microbiol*. **76**:2295–303. https://doi.org/10.1128/AEM.02462-09

91. Cranenburgh RM, Lewis KS, Hanak JAJ. (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

92. Ohashi-Kunihiro S, Hagiwara H, Yohda M, Masaki H, Machida M. (2006). Construction of a positive selection marker by a lethal gene with the amber stop codon(s) regulator. *Biosci Biotechnol Biochem*. **70**:119–25. https://doi.org/10.1271/bbb.70.119

93. Rosano GL, Morales ES, Ceccarelli EA. (2019). New tools for recombinant protein production in Escherichia coli: A 5-year update. *Protein Sci*. **28**:1412–22. https://doi.org/10.1002/pro.3668

94. Dumon-Seignovert L, Cariot G, 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**:203–6. https://doi.org/10.1016/j.pep.2004.04.025

95. Vijayendran C, Polen T, Wendisch VF, Friehs K, Niehaus K, 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**:747–61. https://doi.org/10.1016/j.jbiotec.2006.12.026

96. Marisch K, Bayer K, Scharl T, Mairhofer J, Krempl PM, 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**. https://doi.org/10.1371/journal.pone.0070516

Figure 1

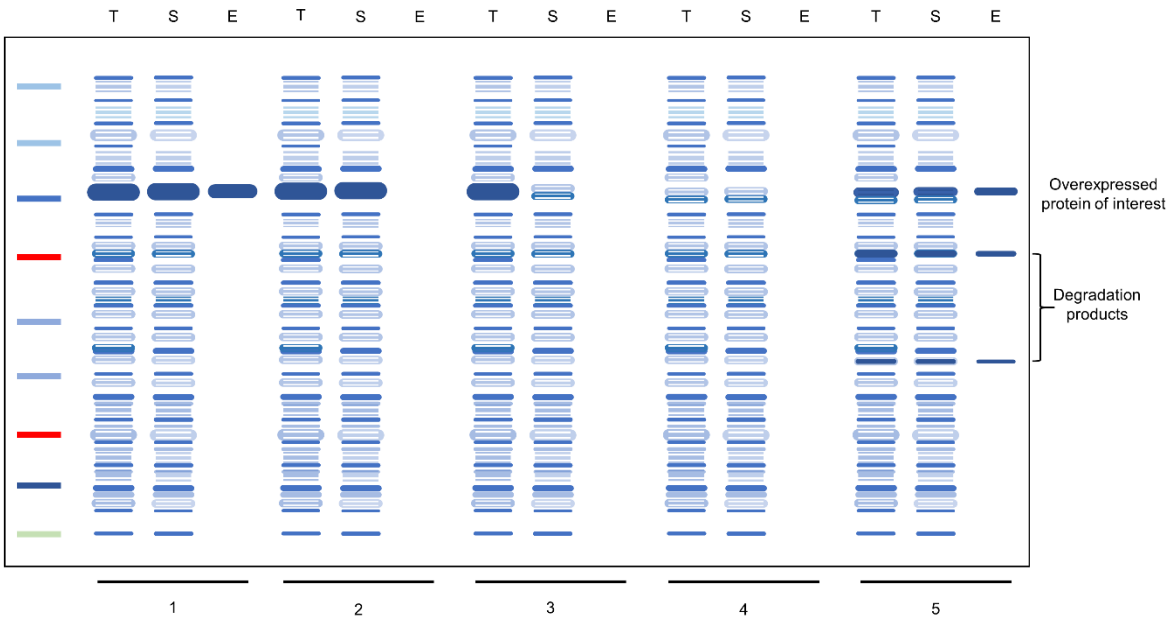

T S E T S E T S E T S E T S E
Overexpressed protein of interest
Degradation products
1
2
3
4
5

Figure 2

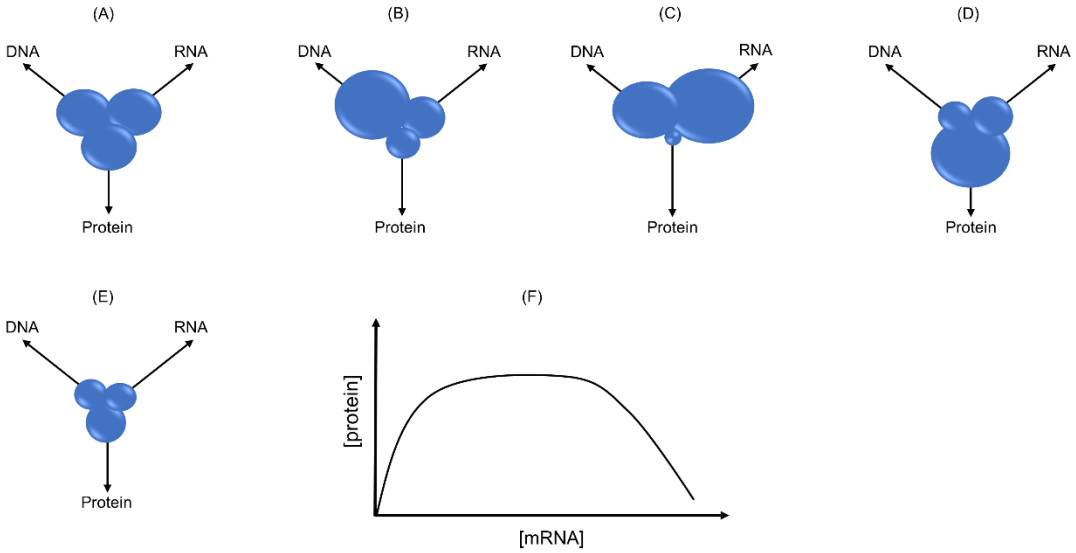

(A)
DNA
RNA
Protein
(B)
DNA
RNA
Protein
(C)
DNA
RNA
Protein
(D)
DNA
RNA
Protein
(E)
DNA
RNA
Protein
(F)
[protein]
[mRNA]

Figure 3

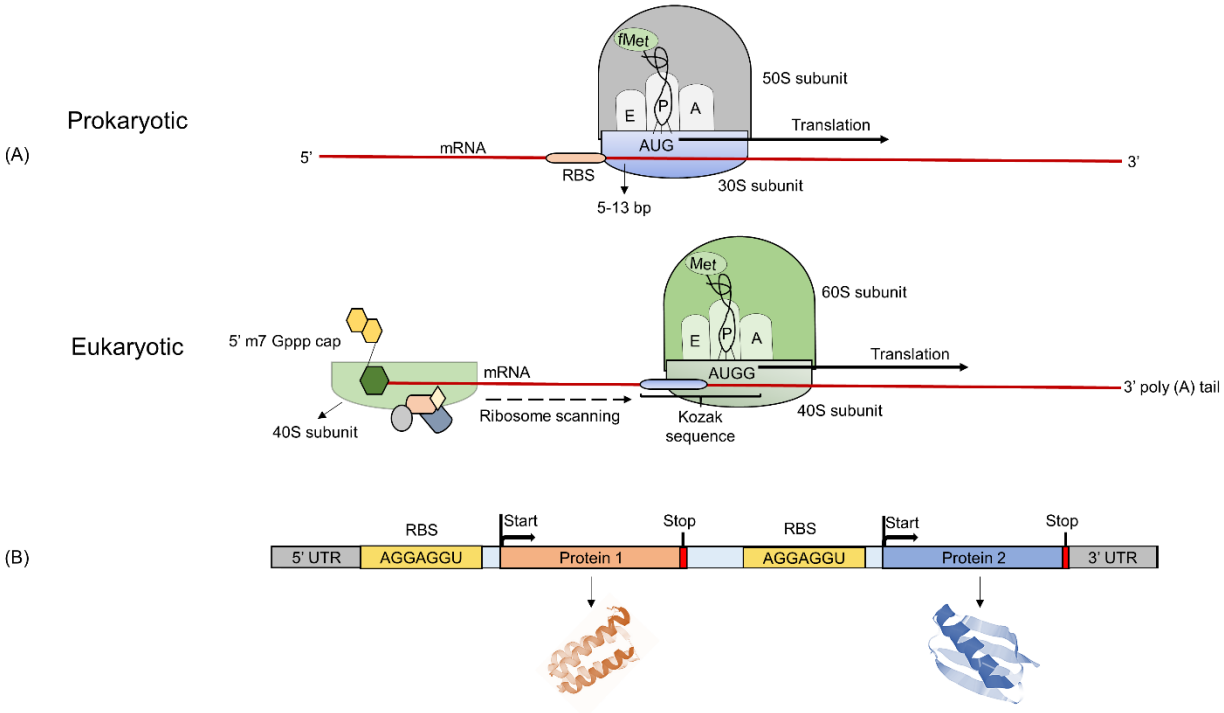
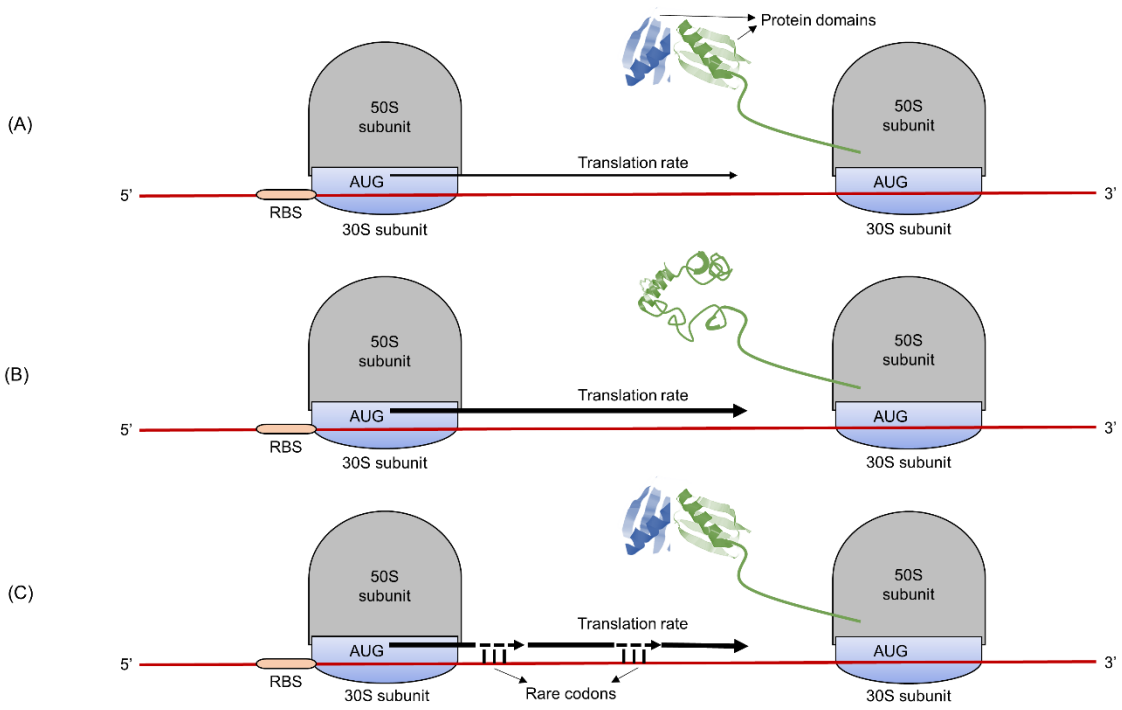

(A)
Prokaryotic
fMet
50S subunit
E
P
A
Translation
5'
mRNA
AUG
3'
RBS
30S subunit
5-13 bp
Eukaryotic
Met
60S subunit
5' m7 Gppp cap
E
P
A
Translation
mRNA
AUGG
3' poly (A) tail
40S subunit
Ribosome scanning
Kozak sequence
40S subunit
(B)
RBS
Start
Stop
RBS
Start
Stop
5' UTR
AGGAGGU
Protein 1
AGGAGGU
Protein 2
3' UTR

Figure 4


(A)
Protein domains
50S subunit
AUG
Translation rate
5'
RBS
30S subunit
3'
(B)
(C)
Rare codons