Engineering microbial cell factories for protein production

Kiavash Mirzadeh

Academic dissertation for the Degree of Doctor of Philosophy in Biochemistry at Stockholm University to be publicly defended on Friday 21 September 2018 at 10.00 in Magnélisalen, Kemiska övningslaboratoriet, Svante Arrhenius väg 16B.

Abstract

Proteins are often produced using microbial cell factories for academic or industrial purposes. Protein production is however not an open-and-shut procedure. Production yields often vary in an unpredictable and context dependent manner, limiting the rational design of a straightforward production experiment.

This thesis gives an overview of how proteins are biosynthesised in bacterial cells and how this knowledge is used to produce proteins recombinantly in a host organism such as *Escherichia coli*. In the present investigation, we reason that unpredictable and poor protein production yields could result from incompatibility between the vector derived 5' UTR and the 5' end of the cloned CDS which leads to an unevolved translation initiation region (TIR). Data presented in this thesis show that an unevolved TIR could work more efficiently and yield more produced protein if subjected to synthetic evolution. Clones with an engineered synthetically evolved TIR showed enhanced protein production in both small- and large-scale production setups. This engineering method could lower production expenses, which in turn would result in increased functional determination of proteins and expanded availability of protein-based medicine to people globally.

Keywords: *Protein production, expression vector, recombinant DNA, Translation initiation region, Escherichia coli, mRNA secondary structure, Synthetic evolution.*

Stockholm 2018
http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-158482

ISBN print: 978-91-7797-366-9
ISBN PDF: 978-91-7797-367-6

Department of Biochemistry and Biophysics

Stockholm University, 106 91 Stockholm

ENGINEERING MICROBIAL CELL FACTORIES FOR PROTEIN PRODUCTION

Kiavash Mirzadeh

Stockholm
University

Engineering microbial cell factories for protein production

Kiavash Mirzadeh

ISBN print: 978-91-7797-366-9
ISBN PDF: 978-91-7797-367-6

Printed in Sweden by Universitetsservice US-AB, Stockholm 2018

To my parents

List of publications

I. **Enhanced Protein Production in *Escherichia coli* by Optimization of Cloning Scars at the Vector-Coding Sequence Junction**
Mirzadeh K, Martíez V, Toddo S, Guntur S, Herrgård MJ, Elofsson A, Nørholm MM, Daley DO.
ACS Synth Biol. (2015) 18;4(9):959-65.

II. **Codon Optimizing for Increased Membrane Protein Production: A Minimalist Approach**
Mirzadeh K, Toddo S, Nørholm MM, Daley DO.
Methods Mol Biol. (2016) 1432:53-61.

III. **TARSyn: Tunable Antibiotic Resistance Devices Enabling Bacterial Synthetic Evolution and Protein Production.**
Rennig M, Martíez V, Mirzadeh K, Dunås F, Röjsäter B, Daley DO, Nørholm MM.
ACS Synth Biol. (2018) 16;7(2):432-442.

IV. **Synthetically evolving translation initiation regions for protein production**
Shilling P*, Mirzadeh K*, Elfageih R, Koeck Z, Rennig M, Nørholm MM, Daley DO.
Manuscript.

* These authors contributed equally.

Additional publications

V. Coordinated disassembly of the divisome complex in *Escherichia coli*
Söderström B, Mirzadeh K, Toddo S, von Heijne G, Skoglund U, Daley DO.
Mol Microbiol. (2016) 101(3):425-38.

Abstract

Proteins are often produced using microbial cell factories for academic or industrial purposes. Protein production is however not an open-and-shut procedure. Production yields often vary in an unpredictable and context dependent manner, limiting the rational design of a straightforward production experiment.

This thesis gives an overview of how proteins are biosynthesised in bacterial cells and how this knowledge is used to produce proteins recombinantly in a host organism such as *Escherichia coli*. In the present investigation, we reason that unpredictable and poor protein production yields could result from incompatibility between the vector derived 5' UTR and the 5' end of the cloned CDS which leads to an unevolved translation initiation region (TIR). Data presented in this thesis show that an unevolved TIR could work more efficiently and yield more produced protein if subjected to synthetic evolution. Clones with an engineered synthetically evolved TIR showed enhanced protein production in both small- and large-scale production setups. This engineering method could lower production expenses, which in turn would result in increased functional determination of proteins and expanded availability of protein-based medicine to people globally.

Table of Contents

Abbreviations

E. coli	*Escherichia coli*
CDS	Coding sequence
TIR	Translation initiation region
UTR	Untranslated region
SD	Shine-Dalgarno
DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
tRNA	Transfer ribonucleic acid
mRNA	Messenger ribonucleic acid
ATP	Adenosine triphosphate
GTP	Guanosine triphosphate
LPS	Lipopolysaccharide
IM	Inner membrane
OM	Outer membrane
GFP	Green fluorescent protein
FACS	Fluorescence activated cell sorter
GOI	Gene of interest
OD	Optical density
AU	Arbitrary units
LB	Luria Broth

Proteins and protein-based drugs

Proteins were first described in 1838 [1]. Since then, we have discovered that proteins fulfil essential functions for all forms of life. Understanding the function of these molecules in depth gives us insight into how life has begun and how it has evolved. A deeper understanding also allows for effective pharmaceutical intervention to restore cellular health, as dysfunction or absence of proteins due to alterations in the coding sequence often leads to pathologies. Examples of such mutation-caused diseases include diabetes [2], amyotrophic lateral sclerosis [3] and sudden death among young people due to heart failure [4]. Patients suffering from such diseases are often treated with protein-based drugs for both preventive and therapeutic purposes. Since proteins are such essential molecules, their production and analysis is a major activity for numerous academic, industrial and pharmaceutical research laboratories. Currently, the global market for protein-based drugs exceeds $157 billion per annum [5,6] and is expected to grow ~4% yearly [7]. In addition, industrial enzymes used within our homes (*e.g.* detergents, textile, paper and pulp, and personal care products) as well as in industrial processes (*e.g.* agriculture feeds, enzymes used in brewing, baking and in producing oils and fats) are estimated to have a market value over $4 billion per annum [8].

Since the emergence of biotechnology and recombinant DNA technologies (gene and protein engineering) in the early 1970's, protein production has extensively been carried out using cells as protein factories. Microbial cells such as *Escherichia coli* are frequently used by academic and industrial laboratories to obtain protein of interest rapidly without the need of either natural sources of animal or plant tissue or large volumes of biological body fluids. Although protein production in microbes has been optimised in the past decades, expression yields still vary in an unpredictable and context dependent manner. This impedes the rational design of protein production using recombinant sources and leads to increased costs and complications in subsequent isolation steps. To tackle this problem and gain insight into how proteins can be made in a reliable way, I present a new method for harnessing *E. coli* as an efficient protein production platform. The method enhances translation initiation efficiency, which is the rate-limiting step in protein biosynthesis and ultimately has a direct impact on protein production yields.

This thesis will, in the first half, give a detailed molecular overview of what is known about protein biosynthesis in *E. coli*, an organism extensively used during my PhD studies. How this knowledge is harnessed for recombinant protein production purposes is described in the second half.

Protein synthesis in *E. coli*

In a bacterium like *E. coli*, large macromolecular machines called ribosomes catalyse the synthesis of proteins. Ribosomes are able to translate the genetic code preserved in the messenger ribonucleic acid (mRNA) into an amino acid polymer in a process termed translation. The ribosomes are made up of both ribosomal RNA (rRNA) and proteins. The *E. coli* ribosome is composed of a large 50S subunit (*circa* 30 proteins, 23S rRNA and 5S rRNA) and a smaller 30S subunit (21 proteins and 16sRNA) that together assemble into the 70S particle [9]. Each 70S complex contains three binding sites for transfer RNA (tRNA) – the molecules harbouring the aminoacyl moieties that become incorporated into the growing polypeptide chain. The binding sites are designated the aminoacyl (A), peptidyl (P) and exit (E) sites. The A-site acts as an entry spot for incoming aminoacyl-tRNA, the P-site holds the elongating nascent polypeptide chain bound to a tRNA whilst uncharged tRNAs get ejected through the E-site. The 30S subunit interacts with the mRNA and the anticodon stem-loop of the tRNA whilst the 50S subunit binds acceptor arms of tRNA and catalyses peptide bond formation between the tRNAs bound to the A and P-sites [10]. The tRNA molecules and the mRNA move concomitantly by the length of three bases, allowing in frame decoding for the next codon [11]. The translation process as a whole can be divided into three major phases titled *initiation*, *elongation* and *termination* (release and recycle) (Figure 1).

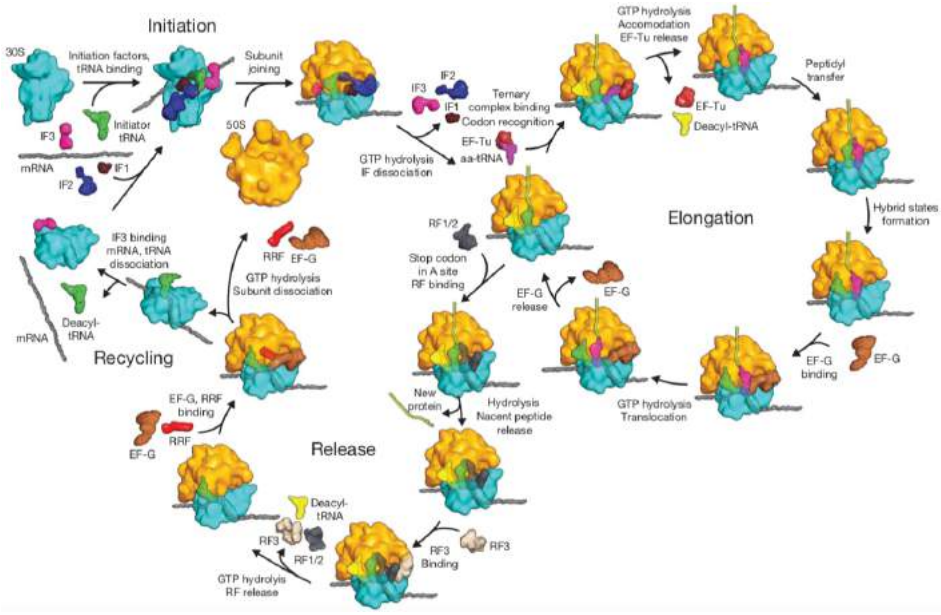

Figure 1. Schematic overview of the main phases in bacterial translation. Details about translation initiation, elongation and termination (release and recycle) are described in the text. Figure taken from [10]. Reprinted with permission.

Translation initiation

Although all steps in translation contribute to protein synthesis, translation initiation is considered to be the rate-limiting step [12,13]. The initial phases of translation initiation begins with the binding of two initiation factors (IF1 and IF3) to the 30S subunit. IF3 prevents premature 70S formation while IF1 blocks the A-site thus preventing entrance of aminoacyl-tRNA. In parallel, the initiator tRNA carrying formylated and aminoacylated methionine (fMet-$tRNA_f^{Met}$) binds to the P-site with assistance from initiation factor IF2. The 30S pre-initiation complex (30S PIC) carrying the initiation factors, mRNA and fMet-$tRNA_f^{Met}$ is then rearranged and stabilised in a way that favours AUG codon-CAU anticodon complementary interactions between the mRNA and the fMet-$tRNA_f^{Met}$. Such rearrangements generate the 30S initiation complex (30S IC). Upon formation of the 30S IC, IF1 and IF3 are ejected, whereas IF2 is ejected after 50S subunit docking on the 30S IC, generating a 70S IC set for dipeptide formation [14]. These initial phases of translation require energy derived from hydrolysis of guanosine-triphosphate (GTP) which is bound to the initiation factor IF2 [14,15]. The order in which fMet-$tRNA_f^{Met}$ and mRNA interact with the 30S subunit is still unclear, but reports suggest that binding occurs stochastically [16]. The newly formed 70S IC holding a fMet-$tRNA_f^{Met}$ in the peptidyltransferase centre (located on 50S subuinit) is ready for the elongation phase of translation.

Translation initiation region

E. coli expresses *circa* 4200 genes [17]. The translation initiation region (TIR) localised towards the 5' end of each CDS differs in sequence and structure for each gene. Per definition, the mRNA region that binds to the 30S subunit during the early events of translation initiation is defined as the TIR. The TIR covers roughly 15 nucleotides on either side of the start codon [18,19]. Therefore, the number of possible TIR nucleotide combinations are up to 4^{30}, *i.e. circa* a quintillion (10^{18}) permutations. Transcriptome-wide experiments have revealed that mRNA are often folded into complex structures *in vivo* [20,21]. The nucleotide sequence constituting the TIR has, however, during the course of evolution been selected to have a relaxed mRNA structure around the AUG start codon [22,23]. Most likely, such a relaxed structure facilitates ribosome-mRNA interactions during the initial phases of translation, with the overall effect contributing to maintenance of cellular fitness [24]. Based on this evolutionary selection, one could hypothesise that some information about translation efficiency has been embedded in this region of the mRNA.

The mRNA together with tRNA, the 30S subunit and three IFs play a critical role in protein synthesis during the initial phases of translation as described previously. Interestingly though, all factors involved in translation initiation remain constant except the mRNA, which is the only variable differing in sequence and structure within the TIR. This region contains the *Shine-Dalgarno* (SD), *spacer* and the *first six codons*, which together are considered to be influential for translation initiation (Figure 2).

The SD forms complementary interactions with the 16s rRNA and is considered to be a major determinant for translation efficiency in *E. coli* [25]. The length of the purine-rich SD sequence can vary between genes and bacterial species [26,27]. The core consensus 5'-AGGA-3' sequence is however conserved in *E. coli* transcripts, as it forms hydrogen bonds with the 3'-AUUCCUCCA-5' bases located on the 3' end of 16s rRNA [28]. Nucleotide modifications affecting this interaction have been reported to influence gene expression levels by several orders of magnitude [29–32]. The distance separating the AUG start codon and the SD is called the spacer region. The number of nucleotides separating the start codon and SD differ across mRNA transcripts, however 9 nt spacing is most frequently found in *E. coli* [26]. Although the spacer region is often left unmodified during overexpression experiments, it has been shown to have a significant influence on expression, both in terms of optimal distance [33] and composition [34–36]. Most likely, nucleotide modifications to the spacer region alter the 5' mRNA secondary structure and stability which in turn affects the efficiency of translation initiation [23,37,38]. Work presented in this thesis has investigated the influence of the spacer region in expression vectors and supports the mentioned postulations that the mRNA secondary structure and stability are determinants for expression efficiency (see paper I). Posterior to the spacer region is the actual CDS to be expressed. The CDS has a start codon at its proximal 5' end, and the most efficient and representative bacterial start codon is the AUG triplet, coding for methionine [39,40]. However, the efficiency of ribosomal selection of mRNA with different initiation triplets has been reported to be coupled to temperature changes [41]. The region coding for N-terminal amino acids following the start codon is also an integral part of the TIR and several reports, including studies in this thesis (paper II), have shown that codon changes immediately downstream of the start codon

can influence expression levels [37,42–44]. Nucleotide changes in the coding sequence (CDS) could both alter the secondary structure of the TIR and/or change initial elongation rates. Interestingly though, rare codons are enriched near the N-terminus of genes [23,45], and the reason for this has been correlated with reduced mRNA structure around the translation start site and not codon rarity itself [46].

Another region believed to be important for translation initiation is the enhancer region, an adenine and uracil (AU) rich sequences usually found in the untranslated region (UTR) upstream of the SD and therefore outside the TIR. This sequence is believed to be recognised by the ribosomal protein SI during early phases of translation initiation [47,48] and several studies have shown that the presence of such an AU-rich sequence in the UTR increases translation efficiency and yields elevated protein levels [26,49,50].

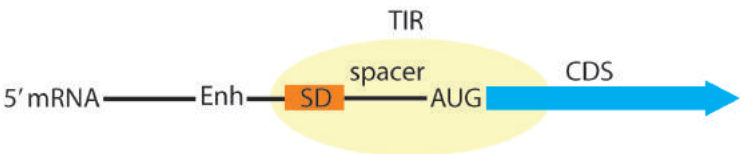

Figure 2. An illustration of the translation initiation region (TIR) together with its elements localised on the 5' end of a bacterial mRNA. The ribosomal footprint (*i.e.* TIR) covers *circa* 15 nucleotides on either side of the AUG start codon (yellow ovoid). The SD, spacer and the outmost 5' CDS are all part of the TIR and influential for translation initiation.

Translation elongation

The formation of a 70S IC with a P-site bound initiator tRNA is required for initiating the elongation phase during translation. Once the 70S IC is formed, incoming charged aminoacylated-tRNA encoded by the second codon binds to the vacant A-site together with an elongation factor EF-Tu. EF-Tu contains a nucleotide-binding site for GTP which upon its hydrolysis leads to the donation of the methionine from the initiator tRNA to the α-amino group of the aminoacyl-tRNA encoded by the second mRNA codon. At this point, EF-Tu and GDP are released. This reaction is catalysed by the 23S rRNA and results in a dipeptidyl-tRNA in the A site and a deacetylated-tRNA in the P site. Upon this, the deacylated-tRNA tilts in a way so it spans the P/E-sites while the dipeptidyl-tRNA tilts so that it spans the A/P sites. Hydrolysis of GTP, and catalysis promoted by EF-G provide the energy required to completely move deacetylated-tRNA and dipeptidyl-tRNA to the E- and P-site respectively while the mRNA is moved with respect to the 30S subunit [10]. This leaves the A-site unoccupied for a new incoming aminoacyl-tRNA to extend the polypeptide during a new round of elongation.

Translation termination

The elongation phase continues until a stop codon is encountered within the A-site. Unlike decoding of sense codons, stop-codon recognition relies on release factors (RFs) which, in contrast to tRNAs, trigger the release of the polypeptide chain. There are two classes (1 and 2) of RFs in bacteria, and their recruitment depends on the nature of the stop codon

triplet. UAG and UGA stop codons are recognised by RF1 and RF2 respectively whilst the UAA stop codon can be recognised by both RFs. The binding of RF 1 or 2 to a stop codon in the A-site leads to the hydrolysis and release of the polypeptide chain from the tRNA. Once the polypeptide chain has been released, RF-3 binds and promotes rapid release of deacyl-tRNA and RF 1 or 2 at the expense of GTP hydrolysis. Then, the 70S complex is disassembled into its smaller subunits through a ribosomal recycling factor, EF-G via GTP hydrolysis [51]. At this point, the mRNA is released and can undergo a new round of translation. It has been estimated that a single amino acid addition to the growing polypeptide chain requires four GTP molecules [52] making protein synthesis energy demanding. Protein biosynthesis is therefore tightly regulated and responsive to different conditions.

A deeper understanding of the underlying principles that govern the efficiency of translation events are highly coveted. Attempts of making it more efficient could have a large overall impact, not least for large-scale, energy demanding protein production setups.

Protein trafficking in *E. coli*

In order to fulfil their function, proteins need to be translocated to their final destination. Bacterial proteins reside in four different compartments (in gram-negative bacteria such as *E. coli*) (Figure 3). These four compartments are known as the cytoplasm, the inner membrane (IM), the periplasm and the outer membrane (OM). The three latter compartments form a cell envelope to encapsulate and protect the cytoplasm, where most of the proteins are localised together with the genome. The cell envelope is composed of three main layers in *E. coli* that act as a barrier against potentially hostile extra-cellular molecules [53]. The IM provides an impenetrable innermost barrier to polar compounds and molecules, consisting of a symmetric bilayer of phospholipids. Notably, approximately 1000 out of 4288 gene products (20-30 %) in *E. coli* are predicted to fully or partly reside in the IM [54]. These proteins are labelled IM proteins and are involved in essential cellular processes such as cell division, cellular respiration, and efflux and influx transportation and represent the largest mass portion of the IM [55]. On the exterior side of the IM, a scaffold polymer of amino acids and sugars make up the peptidoglycan layer which together with proteins provides rigidity, cell shape and protection against osmotic pressure [56]. The peptidoglycan is mainly structured by linear glycan strands cross-linked to peptides. This structural feature exist in all bacterial organisms, although slight variations occur in the fine structure depending on the growth phase, growth medium and presence of antibiotics. The periplasmic proteome contains *circa* 350 proteins [57], including a cluster of proteins called the penicillin-binding proteins (PBPs) that play a key role in the biogenesis of the peptidoglycan layer. Therefore, these proteins have been targets of various antibiotics that aim to inhibit cell-wall synthesis and promote

cell death [58]. Importantly, the environment in the periplasm is oxidising (in contrast to the reducing cytoplasm) thus enabling proper folding of proteins that require disulphide-bond formation [59]. The periplasm is also an environment rich in chaperones and with relatively low protease levels, making it a suitable compartment for producing proteins with high solubility and stability. The outermost barrier of the cell is known as the OM. It consists of two leaflets, with the inner leaflet containing phospholipids and the outer leaflet comprised of lipopolysaccharides (LPS). The extent of permeability is different between the OM and the IM. This is mainly due to existence and structure of the proteins located in the OM. The presence of barrel-shaped proteins (*i.e.* porins) facilitates diffusion of small non-polar and polar molecules [60].

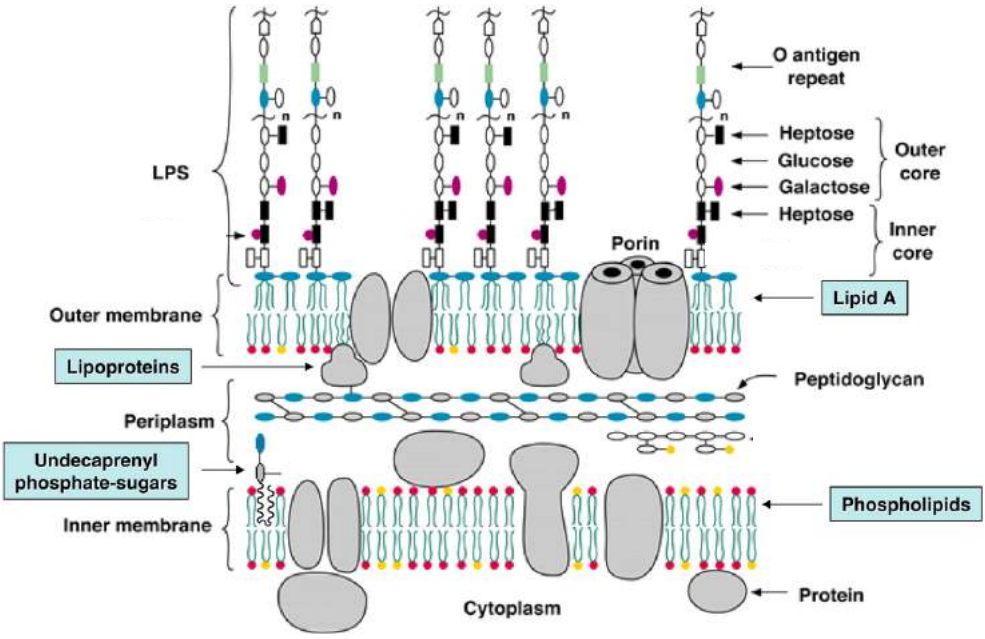

Figure 3. Schematic representation of the four different compartments in *E. coli* including the inner and outer membranes. In *E. coli*, proteins (depicted in grey) perform their function in the cytoplasm, IM, periplasm or the OM. The IM contains IM proteins and separates the cytoplasm from the periplasmic space. The periplasm includes a protein and a peptidoglycan layer that via lipoproteins is bridged to the OM whilst the OM contains OM proteinss and LPS. Figure adapted from [117]. Reprinted with permission.

Protein insertion and translocation across the inner membrane

Proteins residing in the IM, periplasmic space or the OM are actively translocated to their final destination. Insertion into and translocation across the IM is mainly mediated by the Sec-translocon, a hetero-oligomeric protein complex that forms a protein-conducting channel. Three membrane proteins namely, SecY, SecE and SecG form the protein-conduction channel whilst SecA associates peripherally with the SecYEG complex on the cytoplasmic side [61]. Interestingly, SecY and SecE are conserved and essential across all kingdoms of life, suggesting that the principal events during protein translocation occur similarly in all organisms. Generally, proteins localised in the IM are integrated into the IM through the Sec-translocon co-translationally, (*i.e.* whilst nascent polypeptide synthesis is occurring). In contrast, proteins localised in the periplasm or the OM are translocated via the Sec-translocon post-translationally (*i.e*. once the nascent chain has been synthesised and released from the ribosome) [62,63] (Figure 4). Information about the route integral IM and secretory proteins take is embedded in the N-terminal portion of the nascent chain [64]. N-termini containing both positively charged and hydrophobic amino acids plus a cleavage site (signal sequence) are destined for secretion. In contrast, non-cleavable and more hydrophobic N-termini portions (signal anchor sequences) dictate IM integration [65,66]. The average size of a signal peptide is 20 amino acids and the degree of its hydrophobicity is a major determinant for whether the protein will be co- or post-translationally inserted/translocated [67].

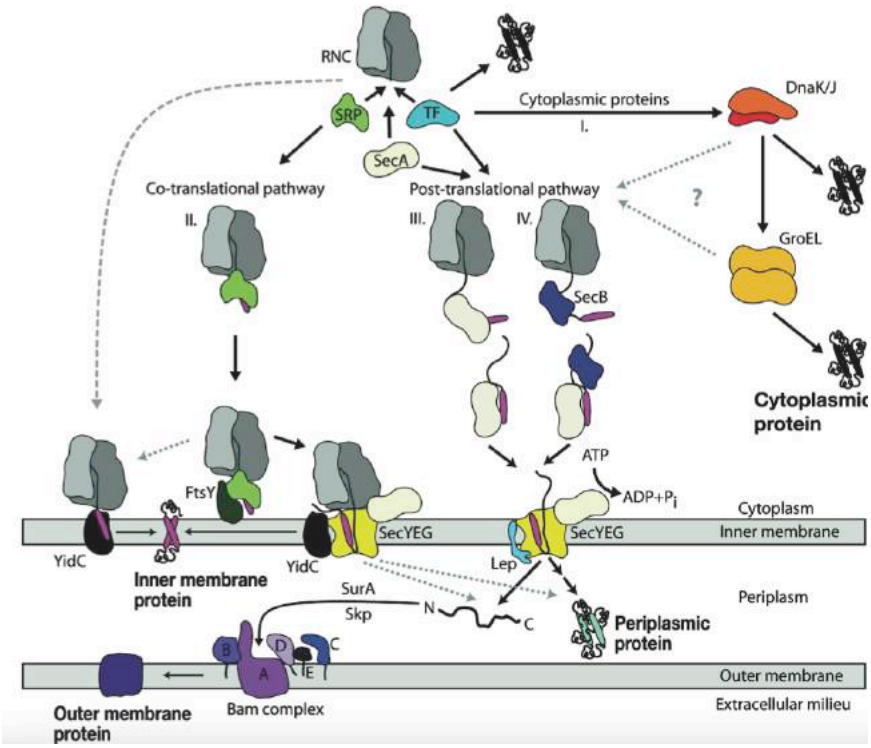

Figure 4. A schematic overview of protein trafficking in *E. coli*. Protein production is always initiated and carried out by ribosomes in the cytoplasm. Proteins residing in the inner membrane (IM) are trafficked via the co-translational pathway, whilst proteins destined for the periplasm or the outer membrane (OM) are secreted through the IM via the post-translational pathway. Translocation machineries such as the Sec translocon and the BAM complex facilitate IM and OM protein insertion respectively whilst chaperones in the cytoplasm and periplasm facilitate protein folding in both compartments (*e.g.* DnaK/J, GroEL and SurA/Skp). The events illustrated in this figure are described in the thesis. Figure source [116]. Reprinted with permission.

Co- and post-translational pathways

Co-translational insertion is mediated by the signal recognition particle (SRP), which identifies signal anchor sequences as they emerge from the ribosome and arrests translation so that premature folding is prevented until the ribosome-nascent chain complex reaches the IM [62]. The ribosome together with its nascent chain are then targeted to a membrane bound SRP receptor called FtsY [68,69]. Conformational changes release both SRP and its receptor from the ribosome once the nascent chain reaches the SecYEG complex. This process is driven by GTP hydrolysis, and both the SRP and FtsY contain nucleotide-binding sites [70]. Thus most IM proteins are inserted into the IM concurrent with protein translation.

In contrast to IM proteins, most periplasmic and OM proteins are translocated post-translationally [63]. In these events, a ribosomal-bound chaperone called Trigger factor (TF) binds to the less hydrophobic signal peptides and translation continues without ribosomal docking to the SecYEG complex [71]. The cytoplasmic chaperone SecB keeps the mature parts of the emerging nascent chain in an unfolded and translocation-competent state [72]. SecB then delivers the unfolded protein to its receptor SecA which in turn is associated to the SecYEG complex [73]. The membrane-bound ATPase SecA is then able to push the nascent chain across the protein-conducting channel through ATP-hydrolysis. Although it is believed that the essential energy requirements for SecA–dependent protein translocation are achieved through ATP-hydrolysis [74], it has also been shown to be stimulated by the proton motive force [75,76].

Research presented in this thesis has dealt with protein biogenesis with the overall aim being the development of a new strategy to enhance recombinant protein production. The data indicates that translation initiation can be a rate-limiting step for production of proteins in all the compartments. By subjecting the TIRs encoding recombinant proteins to synthetic evolution, TIR variants that elevate protein production levels can be selected and studied.

Protein folding

Proteins need to fold into a three-dimensional conformation to obtain their function. It is widely believed that many polypeptide chains emerging from the ribosome start to co-translationally fold [77,78], even within the ribosomal exit tunnel [79]. Co-translational folding is believed to have evolved to counter protein misfolding [80,81], aggregation [82] (*i.e.* exposure of hydrophobic residues to the aqueous environment) and degradation [83]. To facilitate proper protein folding, a set of proteins called chaperones function as folding assistants. Most likely, the first chaperone that most cytoplasmic nascent chains encounter during protein synthesis is TF, since it associates with the ribosome at the exit tunnel [84]. If a protein requires further assistance to reach its native form, it is transferred to downstream chaperones such as DnaK/DnaJ and GroEL/ES [85]. The latter has a cylindrical barrel (GroEL) with a lid (GroES) like structure that can encapsulate proteins and subject them to several rounds of unfolding and refolding until the correct fold is reached [86]. Interestingly, this chaperone system is essential for cell viability, suggesting that other essential proteins require its assistance.

Microbial organisms as cell factories

Bacterial and yeast host organisms are frequently used as main protein production platforms and examples of such hosts include *Escherichia coli*, *Bacillus subtilis* and *Saccharamyces cerevisiae*. The choice of a microbial host organism for protein production is usually based on the origin, characteristics and compartmental residence of the recombinant protein. For instance, proteins that require post-translational modifications (*e.g.* disulphide-bond formation, glycosylation and subunit assembly) are usually produced in *S. cerevisiae*, which have the advantage of being unicellular (*i.e.* easy to genetically manipulate and grow fast) but yet have protein-processing capabilities similar to higher eukaryotic organisms. *S. cerevisiae* is also extensively studied as a eukaryotic model organism, which allows researchers to use available knowledge to further engineer the organism for protein production proposes. However, one drawback with proteins produced in yeast organisms is that they contain many N-glycosylations that differ in structure compared to human-like glycosylations. This could affect the produced protein's half-life, as it could be recognised as foreign by the host immune system and therefore reduce its effectiveness as a therapeutic drug [87].

Another microbial host often used for extra-cellular secretion of recombinant protein is *B. subtilis*, a gram-positive bacterium found in the upper layers of soil. *B. subtilis* facilitates protein secretion more efficiently compared to other bacterial host organisms because its cell is only composed of one membrane. In addition, industrially produced enzymes

used in food or washing detergents are mainly produced in *B. subtilis* due to the fact that it is non-pathogenic. Heterologous genes are also more likely to be successfully expressed in *B. subtilis* because of its unbiased nucleotide composition. Although it is heavily used in industrial protein production settings, genetic tools for its strain manipulation are still lacking compared to other frequently used organisms [88].

E. coli as a protein production platform

As well as being a model organism for understanding basic molecular events of life, *E. coli* has become a well-established and widely used cell factory for cost-effective protein production. This is mainly due to its well-understood physiology, fast growth rate and the fact that it is easy to manipulate [89]. Most likely, it is among the most thoroughly studied organisms to date. Historically, it has been a workhorse organism that has paved the way for commercialising the first biopharmaceuticals. For example, synthetic human insulin became the first approved genetically-manipulated drug to be produced in *E. coli* [90]. In addition, *E. coli* is currently used to produce *circa* 30% of all approved protein-based pharmaceuticals [91,92].

Given optimal conditions, *E. coli* has a doubling time as short as 20 minutes. However, it should be noted that induction of the gene of interest and the presence of antibiotics in the growth culture might impose a cellular burden, which in turn decreases the doubling time. Such metabolic stress also decreases biomass formation over time, resulting in hampered protein production yields [93,94]. Other factors that might make expression levels difficult to predict *a priori* include (1) non-familiar

nucleotide composition of the gene sequence (*e.g.* GC-rich/poor regions) [95], (2) misfolding of the protein leading to inclusion bodies [96], (3) incompatibility between the host's codon usage and the gene codon sequence [97], (4) toxicity of the protein that leads to degradation by proteases and/or cell death [98] and (5) stability or instability of certain regions in mRNA affecting translational and mRNA decay [19]. Genetically altered *E. coli* strains with different characteristics have been developed to circumvent some of these difficulties that might be encountered during a protein production experiment. Examples of such *E. coli* strains include the BL21 strain series, which range from allowing tightly controlled expression to overcoming the effect of codon bias and overexpression of toxic and membrane proteins [99] (Table 1).

E. coli strain	Used for	Characteristics at genetic level
ArcticExpress	Grow at low temperature	Low temperature adapted chaperone encoded gene
BL21	Less protease degradation of recombinant protein	*Lon* and *OmpT* protease deficient
BL21-Codonplus (RIL)	Overcome the effect of codon biasness (AT rich gene)	tRNA encoding gene (*argU, ileY* and *leuW*)
BL21(DE3)	Expression under T7 promoter	T7 pol encoded
BL21(DE3)pLys S/E	Controlled expression	Lysozyme encoded plasmid
BL21 Star	Increase stability of mRNA	*rne31* gene mutated
C41(DE3)	Overexpressing toxic proteins	Carry the lambda DE3 lysogen which expresses T7 RNA polymerase from the lacUV5 promoter by IPTG induction
C43(DE3)	Membrane and globular protein	-Same-
Codon plus (RP)	Overcome the effect of codon biasness (GC rich gene)	tRNA encoding gene (*argU* and *proL*)
Lemo21(DE3)	Membrane, globular and toxic protein expression	Lysozyme encoded under rhamnose inducible promoter
M15	Gene under T5 promoter	Constitutively expresses *lac* repressor at high levels
Origami	Protein required disulphide bond formation	*gor* and *trxB* genes mutated
Rossetta	Both the AT and GC rich gene	All the rare tRNA coding gene
SG13009	Enabling *trans* repression of protein expression prior to IPTG induction	Carry the repressor plasmid pREP4, which constitutively expresses *lac* repressor at high levels
Shuffle	Proper disulphide bond formation	*DsbC* gene encoded

Table 1. Commonly used *E. coli* strains for recombinant protein production. Table taken from [99]. Reprinted with permission.

To equip cell factories with large amounts of mRNA transcripts that code for the gene of interest, the BL21 (*DE3*) strain with a genomically integrated lysogenized *DE3* phage fragment encoding T7 RNA polymerase has been developed. In the presence of a T7-based promoter, it can transcribe the gene of interest 5-8 times faster than the endogenous *E. coli* RNA polymerase [100], thus ensuring that mRNA levels are not a limiting factor during a protein production experiment. In addition, the BL21 (*DE3*) p*LysS/E* strains have been developed to minimise leaky expression, which is of interest when the gene to be expressed is toxic to the host. This has been done through incorporation of a plasmid-encoded lysozyme that inhibits background-levels of T7 RNA polymerase formation prior to gene induction. Occasionally, expression levels can be low due to differences between the codon usage preference in *E. coli* and the codons present in the gene to be expressed. To tackle this, strains containing plasmids that expand the tRNA pool, thereby compensating for codons rarely used in *E. coli* have been engineered. Examples of such strains include the Rosetta and the CodonPlus-RIL strains, both derivatives of the BL21 (*DE3*) strain [99]. Furthermore, overexpression of membrane or secretory proteins can saturate the secretory machineries embedded in the IM of *E. coli*. To address this issue, strains such as C43/41 (*DE3*) have been isolated. Such strains contain mutations in the *lac*UV5 promoter, which lead to decreased levels of T7 RNA polymerase, ultimately reducing synthesis of the mRNA of interest. This in turn relieves the secretory machinery – a bottleneck for membrane protein production in bacteria [101]. The overexpression of proteins can also cause the formation of insoluble aggregates known as inclusion bodies. In order to decrease inclusion body formation, the incubation temperature of the growth medium is usually decreased post induction. Inclusion bodies can also be solubilised and refolded into active protein using denatur-

ants [102]. Although complex post-translational processes are absent in *E. coli*, researchers have been able to successfully transfer and engineer functional glycosylation pathways used in other organisms into *E. coli* [103]. Such examples, among with other technological advancements, highlight the versatility of *E. coli* and its applicability as a protein production host. Overall, it can be argued that *E. coli* has been evolved as a cell factory to coordinate with protein production needs. However, there is still room for improving it as a host organism, particularly when it comes to bio-sustainable chemical production using metabolic engineering [104].

Vector design for recombinant protein production

The knowledge described in the first half of this thesis is often harnessed to efficiently produce proteins recombinantly in large-scale production set-ups. Recombinant protein production involves engineering and experimental design around systems that allow controlled expression of the gene. The CDS is typically cloned into an expression vector that contains well-defined genetic elements selected for maximum production. Such vectors contain promoter sequences that control transcription and permit high levels of mRNA. Selection of high-copy vectors can also increase gene dosage though their origin of replication. Antibiotic resistance markers are also selected for to ensure the expression vector's intracellular maintenance. In order to enhance mRNA recognition by the ribosome, a strong and optimally positioned SD is chosen. The most commonly used vectors for protein production are the T7-based pET range, which contain all the genetic elements mentioned above. They allow expression of the CDS in strains of *E. coli* that contain

a lysogenized DE3 phage fragment encoding the T7 RNA polymerase in their genome (*e.g.* BL21(DE3) and derivatives that have been selected and/or engineered for high-level production). Finally, growth culture conditions (*e.g.* temperature, pH, nutrients and aeration) are monitored and sampled on a case-to-case basis, once the strain and vectors have been engineered [89,105,106].

Despite clear advances in the field, recombinant protein production is still complex and discouragingly, gene expression levels vary in an un-predictable manner even when powerful genetic modules are used [30,55]. In principle it might be related to low transcription efficiency or the degradation of the formed transcript, as well as ineffective translation of the transcript. In addition, degradation, mis-targeting or incorporation into aggregates or inclusion bodies of the translated protein all influence the steady-state concentration of the protein (Figure 5). This is often the case when IM and secretory proteins are recombinantly expressed, as any attempt to increase their production rate might influence their complex biogenesis pathways and increase the likelihood of mis-targeting and aggregation formation (due to their hydrophobic nature).

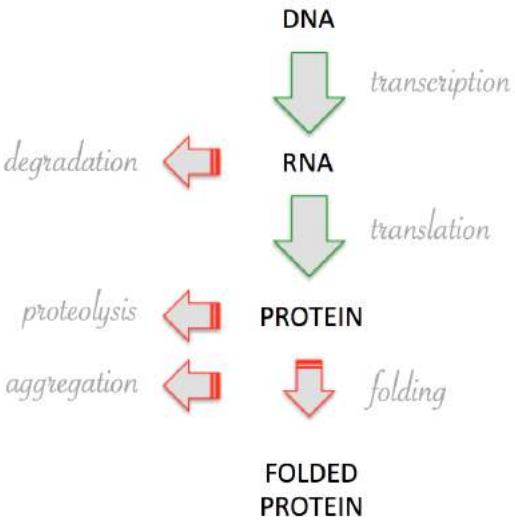

Figure 5. The steady-state concentration of a recombinatly expressed gene product is affected by several processes.

The difficulties of obtaining high yields of membrane proteins is partly reflected in their under-representation in the Protein Data Bank (PDB) as only ~1% of all solved atomic structures belong to membrane proteins. This is not due their lack of natural prevalence nor importance, as 20-30 % of all proteins in both pro- and eukaryotes are membrane embedded [107] whilst 70 % of all drugs act on membrane proteins [108]. As mentioned, one major side-effect of IM overexpression is the formation of aggregated inclusion bodies containing the target IMP together with sequestered chaperones and precursor forms of secretory proteins [109]. In addition, cells forced to overexpress IM proteins tend to increase stress-response promoter activity [110] and decrease cellular respiration [109]. Overexpression of CDSs belonging to soluble proteins might also impose a metabolic burden and alter the physiology of the bacterial cell,

especially if they require post-translational modifications or are heterologous [111].

Obviously, many parameters contribute in causing a context-dependent variation in expression levels. This limits the rational design of recombinant expression experiments, leading to increased costs and halts in downstream purification steps. Research presented in this thesis has attempted to address the context-dependent expression variation and identified one common cause; that being incompatibility between the vector driven 5' UTR and the 5' end of the CDS (see paper I). To be specific, the cloning of a CDS into an expression vector generates a random unevolved TIR that has not been subjected to any evolutionary pressure. Data presented in this thesis indicate that TIRs could work more efficiently, and enhance protein synthesis, if subjected to synthetic evolution. Such synthetically evolved TIRs most likely provide a structurally relaxed 5' mRNA that is not sequestered in a stable mRNA secondary structure (paper I) and can be more readily accessed by the ribosomes during the initial phases of translation (Figure 6).

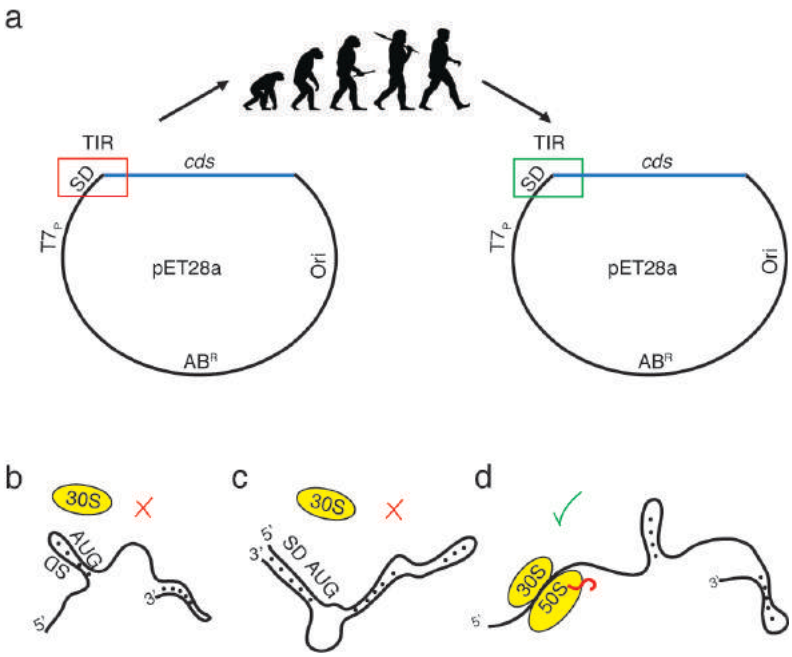

Figure 6. A need to synthetically evolve the TIR for efficient translation initiation and enhanced protein production. (a) After cloning the CDS into an expression vector (*e.g.* pET28a), an unevolved TIR is generated (left panel). This TIR is partly formed from the 5' CDS and partly formed from the vector derived 5' UTR. Such a TIR has not been subjected to any evolutionary pressure like natural TIRs and consequently has an increased likelihood of being sequestered in a double-stranded mRNA either regionally (b) or globally (c) once transcribed. A synthetically evolved TIR (right panel) would on the other hand have an increased likelihood of having a relaxed mRNA structure, facilitating interactions with the ribosome (d) and ultimately lead to enhanced protein production levels.

Summary of papers

The yield of protein production is the sum of several cellular processes such as transcription, translation and protein folding (Figure 5). During my PhD studies, we focused on translation, with the primary goal of better understanding how to make the translation initiation phase more efficient for protein production using *E. coli* as a host organism. Translation initiation is known to be the rate-limiting step in bacterial protein synthesis [12,13], as several key molecules need to assemble with the 30S subunit of the ribosome. The 5' end of the mRNA is however the only variable during the early events of translation and mRNA binding to the 30S ribosomal subunit is the rate limiting step as compared to tRNA and IF binding [112].

Before I started my PhD studies our laboratory had realised that recombinant expression levels vary in an unpredictable way even when powerful genetic elements such as a strong promoter and SD are used [55]. Some years later, it was shown that the stability of mRNA folding near the TIR is a major determinant for gene expression [37,112,113]. Inspired by these observations, former members of our research group showed that selective synonymous codon substitutions immediately downstream of the AUG start codon were more influential for membrane protein production than codon optimising the entire gene [44]. Consequently, one of the major aims during my PhD studies was to better understand how

TIRs could be experimentally engineered for a more predictable protein production.

An interesting observation made during the early stages of my PhD studies was that standard cloning using common restriction sites generates a random, unevolved TIR with an increased likelihood of having its SD sequestered in an mRNA secondary structure (paper I). Such unevolved TIRs were partly formed from the 5' CDS and partly formed from the vector derived 5' UTR and were found to be suboptimal for membrane protein production in *E. coli* (Paper I & II). However, selective mutagenesis on either side of the AUG start codon (*i.e.* the restriction site or the 5' CDS) elevated expression levels (Figure 7). Based on these observations, we realised that there was a need to synthetically evolve the TIR, similar to what nature had done, but in the test tube. The logic behind our hypothesis was that the unevolved TIR generated after cloning is a one in a quintillion permutation and most likely suboptimal for expression. Therefore, the unevolved TIR could work more efficiently during the initial phases of translation if subjected to synthetic evolution. Such a synthetically evolved TIR would have an increased likelihood of having a less stable secondary structure and a non-sequestered SD region (Figure 6d).

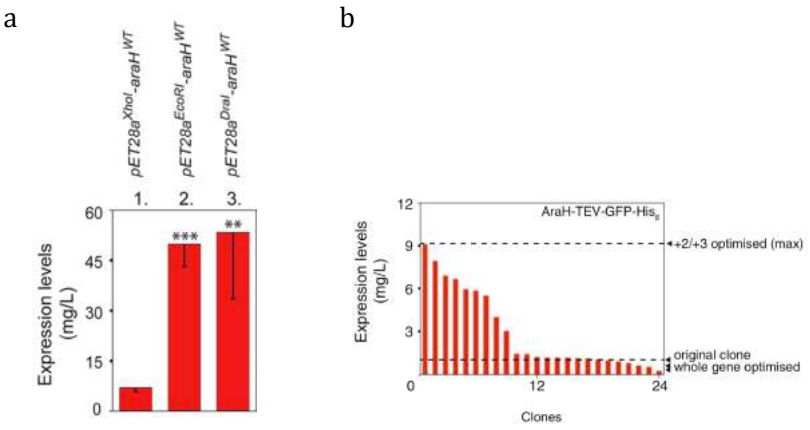

Figure 7. Nucleotide sequences on either side of the AUG start codon influence expression levels for *araH-gfp*. (a) The effect of alterations in the nucleotide composition preceding the AUG start codon. Comparison between clones with a 5' *Xho*I, *Eco*RI or *Dra*I restriction site are shown. (b) A screen of expression levels from 24 clones in a library harbouring all possible combinations of synonymous codons in position +2 and +3 downstream of the AUG start codon. Figure taken from paper I and II. Reprinted with permission.

To obtain a synthetically evolved TIR, we randomised the unevolved TIR by PCR using degenerate primers. These primers partly randomised the nucleotide sequence surrounding the AUG start codon. The six nucleotides upstream of the AUG start codon were fully randomised and the six nucleotides downstream of it were restricted to synonymous codon substitutions only (Figure 8a). The PCR-product was therefore a mixed library of clones ranging from 16 thousand to 50 thousand different TIR variants, depending on the 5' CDS. When we transformed and randomly tested the expression levels of 96 clones from our TIR libraries, we observed 96 unique expression levels (Figure 8b).

a

b

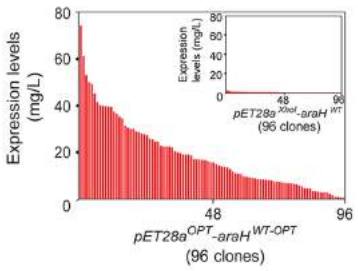

Figure 8. TIR library mutagenesis and screening approach. (a) Schematic representation of mutagenesis approach using degenerate primers that generated TIR libraries. The six nucleotides immediately preceding the start codon were fully randomised (denoted N) and the six nucleotides downstream of the start codon were restricted to synonymous codon changes only (denoted N*). (b) Expression levels of 96 randomly picked clones from the *araH-gfp* TIR library. Inset boxes show cell-to-cell expression variation of 96 clones with an unevlolved TIR. Figure taken from paper I. Reprinted with permission.

To better understand the underlying principles influencing our results, we analysed approximately 1700 TIR variants using FACS and next generation sequencing together with the Nørholm group at Technical University of Denmark (DTU). By assessing the TIR sequences with a computational predictor for mRNA stability (mFold), we could conclude that low GC-content and relaxed mRNA structure around the TIR were important determinants for high production levels (Figure 9). Although the graphed data showed a scattered plot, an average line (representing 100 data points in a sliding window) trended towards a relationship between low mRNA secondary structure around the TIR and high expression level of the CDS. Similarly, when the sequences were analysed using the RBS calculator [29] (a state of the art predictor for translation initiation rates), the trend line was comparable to the one observed using mFold (data not shown).

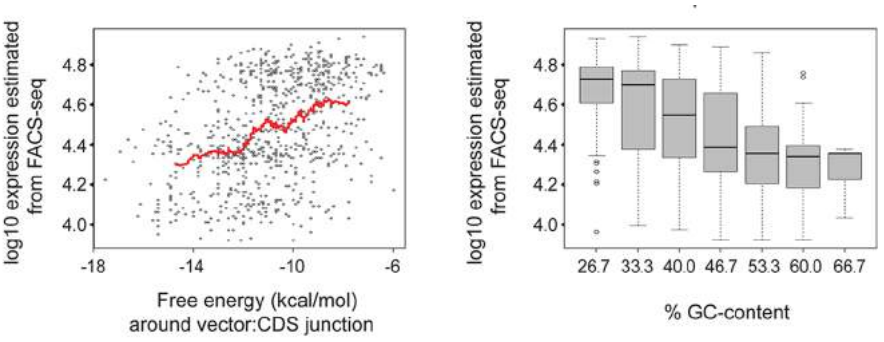

Figure 9. FACS-seq analysis of approximately 1700 TIR variants showed that relaxed mRNA structure (left panel) and low GC-content (right panel) were important determinants for expression levels. Figure taken from paper I. Reprinted with permission.

Although our screening approach allowed detection of clones with enhanced gene expression, it required laborious testing and/or a FACS for deep screening into the libraries. In addition, the expression cassettes yielded a physically fused reporter protein (*i.e.* GFP bound to the protein of interest). In order to tackle this, we devised translational coupling devices together with collaborators in DTU (paper III). Such coupling devices were designed using mRNA secondary structures that sequestered the TIR of an expression reporter placed downstream of the CDS of interest in an operon like manner (Figure 10a). If the upstream CDS was efficiently translated, the helicase activity of the ribosome could untangle the mRNA secondary structure sequestering the TIR of the downstream reporter, enabling *de novo* translation or translation reinitiation at that site (Figure 10b). Consequently, it allowed detection of the translation efficiency of a CDS without creating a physically fused reporter molecule. Moreover, such coupling devices allowed deep screening of our TIR libraries. By sandwiching these devices between the CDS and an antibiotic selection marker, different expression levels in large clone libraries could be screened using a cell survival assay on nutrition plates containing increasing amounts of antibiotics (Figure 10c).

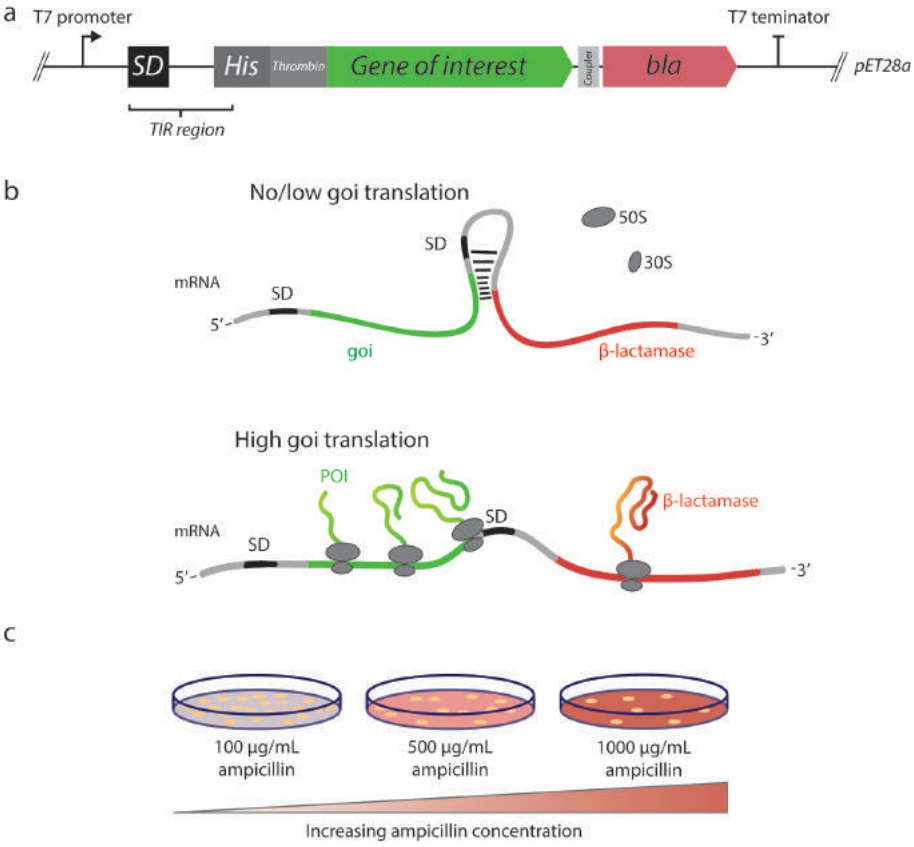

Figure 10. Overview of the expression cassette used and the experimental procedure to obtain clones with a synthetically evolved TIR. (a) Schematic representation of the *gfp* and the *bla* genes sandwiched with a coupling device and cloned into the pET28a vector harbouring a sequence coding for a His_6-Thrombin fragment. (b) Upon induction, no or low levels of translated goi result in low formation of the β-lactamase (upper panel). In contrast, ribosomes successfully translating the first goi untangle the coupling device, which leads to formation of β-lactamase. (c) By plaiting cells harbouring the TIR libraries versus cells transformed with the unevolved TIR onto LB-agar plates containing different concentrations of ampicillin, clones with different translation efficiency could be selected at ampicillin concentrations where the unevolved TIR could not produce enough β-lactamase to sustain growth, but the synthetically evolved TIRs could.

In paper IV, we aimed to better understand which element of the TIR (*i.e.* SD, spacer or CDS) that is most sensitive to nucleotide changes and most amenable for maximum protein production. To map out this region, we generated TIR libraries that vary at the SD, spacer, spacer/CDS and CDS regions (Figure 11) for an expression clone harbouring a genetic setting according to Figure 10a. The expression cassettes contained an N-terminal His_6-tag-Thrombin fragment fused to GFP and a translationally coupled gene encoding for β-lactamase, enabling expression level determination of our TIR libraries using an ampicillin resistance screen (Figure 10c). For each library, a clone with a synthetically evolved TIR, yielding increased protein amounts compared to the unevolved TIR could be selected (Figure 12). Preliminary results indicate that a synthetically evolved TIR from the spacer/CDS library with complete randomisation of the first two amino acids yields most protein for His_6-Thrombin-GFP production. Interestingly, this library had the largest amount of possible TIR permutations compared to the libraries varying the other elements within the TIR. The molecular details behind why the spacer/CDS region is most susceptible to nucleotide changes, and if this region is also most sensitive for more CDSs, will be studied in the near future.

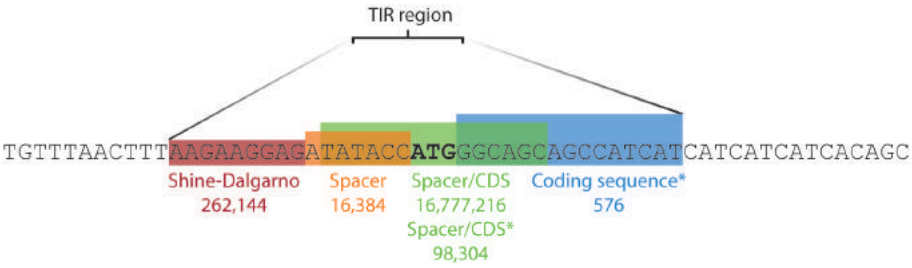

Figure 11. Generation of clone libraries varying in different regions of the TIR. The design enabled complete randomisation of nine nucleotides for the TIR^{SD} and seven nucleotides for the TIR^{SPACER} libraries, complete or partial randomisation of the six nucleotides either side of the AUG start codon for the $TIR^{SPACER/CDS}$ and the $TIR^{SPACER/CDS(*)}$ libraries respectively, and partial randomisation of the first five codons downstream of the AUG start codon for the $TIR^{CDS(*)}$ library. Libraries labelled with an asterisk (*) only allowed synonymous codon changes. The number of possible permutations are indicated below each TIR region.

To date, the best way to modulate translation initiation and protein production is to use *in silico* tools that computationally predict translation initiation efficiency based on thermodynamic principles [19,29,114,115]. To investigate how well experimentally engineered clones with a synthetically evolved TIR perform versus computationally predicted TIRs, we generated and compared the expression of nine additional clones (Figure 12). In every case, the synthetically evolved TIRs engineered in the laboratory yielded more protein compared to the computationally predicted TIRs. Interestingly, only two of the nine computationally predicted TIRs produced more protein than the original unevolved TIR, suggesting that computational predictors may not work in a reliable manner at this point of time.

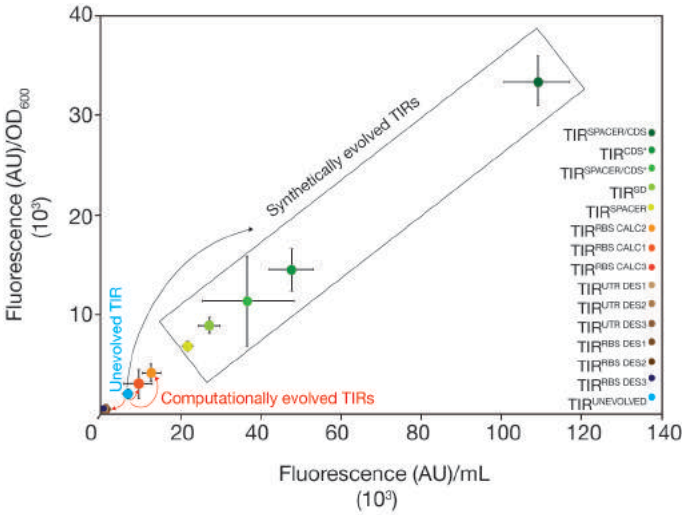

Figure 12. Comparison of *his$_6$-thrombin-gfp* expression between clones mutagenized at different parts of the TIR. Green points represent clones with a synthetically evolved TIR. In parallel, 9 additional computationally predicted TIRs were constructed using three different *in silico* predictors. Expression from all TIR variants are compared to the unevolved TIR (blue mark).

To explore the implications that clones with a synthetically evolved TIR have for large-scale protein production set-ups, we cloned a coupling device cassette into an expression vector harbouring the CDS for an Affibody® molecule. Similarly as described, we were able to select synthetically evolved TIRs that could survive on nutrition plates containing higher levels of ampicillin as compared to the original clone with an unevolved TIR. Ultimately, these synthetically evolved TIRs enhanced the protein production levels in large-scale fermentations experiments (Figure 13). These results indicate that synthetically evolved TIRs could have commercial value, as they further enhanced protein production of an already optimised industrial expression system. Thus, the production levels of other, existing and future protein-based pharmaceuticals could be enhanced by synthetically evolving the TIR.

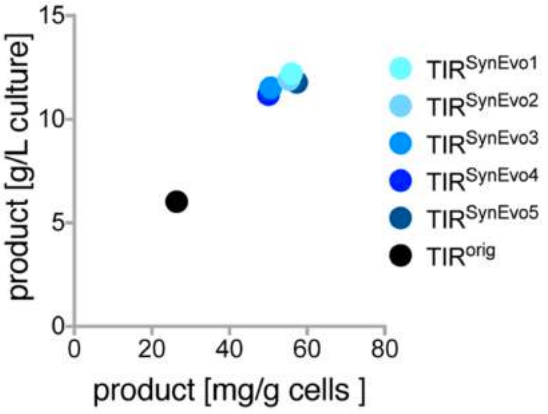

Figure 13. Affibody® production levels in fed-batch fermentation. Production levels were calculated and compared between the original unevolved clone (black) and five different synthetically evolved variants (shades of blue) per cell mass and per culture volume. Figure taken from paper III. Reprinted with permission.

Conclusions and future perspectives

During my PhD studies, we realised that cloning a CDS into any expression vector generates an unevolved, random TIR that is suboptimal for recombinant protein production. Later on, we understood that such an unevolved TIR was one variant out of the quintillion possible permutations; therefore it was likely to be inefficient during the initial phases of translation. This reasoning was made in parallel with the observation of naturally evolved TIRs, which have during the course of evolution been selected for to be more relaxed in this region. Although the intention of such selection is probably coupled to cell fitness [24], we reasoned that that there was a need to synthetically evolve the TIRs for efficient translation initiation and high protein production yields.

Since the synthetically evolved TIRs often tended to have a more relaxed structure and be less GC-rich in general, I envision the development of an *in silico* tool for primer design that would, instead of fully randomising the TIR, predict nucleotide changes with an increased likelihood of having a decreased stability and structure in a certain sequence context. Libraries generated using such primers would then contain clones with less variation but more solutions, increasing the likelihood of identifying a synthetically evolved TIR. In order to reach a state where experiments are aided using computational predictors, certain question marks need to be answered. Firstly, which part of the TIR is most sensitive to nucleotide changes regardless of the CDS context? How much variance is it experimentally possible to generate using degenerate primers? What is

the amount of variance after transformation into the cells? And how large of the variance do we have to experimentally sample to get enough information for reliable computational predictions? These questions could partly be answered using deep sequencing of the libraries prior and post transformation into cells.

The ultimate goal is to computationally design TIRs that could be used with precise and reliable experimental output regardless of the target gene sequence. Today, the experimentally evolved TIRs outperform the computationally designed TIRs when it comes to maximum protein production levels, as shown in this thesis. Perhaps, this is due to a combination of lack of predictive power of the algorithms and a lack of understanding about the underlying principles that govern efficient translation initiation. For instance, the predictors only consider parts of the nucleotide sequence in the vicinity of the TIR, neglecting any global mRNA interactions that might have a large influence on expression levels [20]. Once more detailed insight about translation initiation is gained; *in silico* predictors also become more powerful and vice versa. Therefore, I envision a future were experimental design is at least assisted by computational tools in a reliable and reproducible manner.

Occasionally, we observed that synthetically evolved TIRs worked so efficient that they most likely sequestered a large fraction of the ribosomes from the host cell. This in turn imposed a metabolic load on the cells, which ultimately led to fitness issues and growth impairments over time. Since natural gene expression levels have been evolved to maximise cell fitness during the course of evolution [24], it is likely that the most efficient synthetically evolved TIRs, which have been evolved to maximise protein production cause such a side effect. To solve this, I

think the TIR libraries need to be tested in conjunction with another molecule that reports for cell fitness. Such a reporter molecule (*e.g.* a fluorescent protein) could be genomically integrated and have a promoter sensitive to aggregation formation or cell envelope stress as shown by patent application US2017355983 (A1). By screening the TIR libraries with antibiotic resistance cassettes, one could then distinguish between colonies that report for cell fitness issues and healthy colonies post induction of the gene of interest. Clones with TIR variants that resist the highest amount of antibiotic (*i.e.* contain a synthetically evolved TIR) and have the lowest florescence (reporting for low cellular burden) could be de-convoluted using this approach.

Populärvetenskaplig sammanfattning på svenska

Proteiner är essentiella molekyler som uppfyller nödvändiga cellulära funktioner för alla former av liv på vår planet. En djupare förståelse för hur proteiner fungerar ger oss insikt om hur livet har börjat och utvecklats samt hur ny medicin som återställer eller inhiberar en cellulär process kan konstrueras. För att kunna studera proteiner extensivt, vare sig det är för akademiskt eller industriellt syfte, krävs först stora mängder isolerat protein. Framställningen av proteiner sker vanligtvis i mikrobiella värdorganismer så som *Escherichia coli*. Dessa värdorganismer är utrustade med alla de molekylära verktyg (t.ex. polymeraser och ribosomer) som krävs för en effektiv proteinproduktion.

Translationsinitiering (det initiala steget under proteinsyntes) anses vara hastighetsbegränsande för proteinproduktionsprocessen. Tidigt under mina doktorsstudier upptäckte vi en gemensam orsak för låga proteinproduktionsnivåer; nämligen att kloningen av gener in i DNA-vektorer resulterar i en suboptimal translationsinitieringsregion. Denna region kunde fungera mer effektivt under translationsinitieringen om den utsattes för syntetisk utveckling. Syftet med denna doktorsavhandling var således att utveckla en ny metod som syntetiskt utvecklar translationsinitieringsregion vilket slutligen ökar den totala proteinproduktionen i *E. coli*. Den presenterade metoden kunde appliceras effektivt för både småskaliga och storskaliga produktionsuppställningar. Denna metod kan sänka produktionskostnader, vilket i sin tur skulle

kunna resultera i en ökad förståelse för hur proteiner med okänd funktion fungerar och en ökad tillgänglighet av proteinbaserade läkemedel till fler människor.

Acknowledgements

I will always be grateful to my supervisor **Daniel Daley** for letting me educate myself in his lab, passing on his knowledge to me and improving me as a scientist. I feel lucky to have worked with you and highly appreciate your positive energy and attitude towards science and life. I truly enjoyed my time as a PhD student, five memorable years which I will carry and remember with great pleasure.

Special thanks goes to my present and former group members, in particular **Patrick, Rageia, Claudio, Jörg , Stephen** and **Bill**. Thanks for being great colleagues and making me feel at home at DBB.

Thanks also to our master students **Suchithra**, **Aurelie**, **Zoe** and **James.**

I am also grateful to **Morten, Maja** and **Virgina** for being great collaborators at DTU.

It was a pleasure to perform teaching together with **Johan B**, **Braulio** and **Rickard**.

A big thanks to **Pia** and **Stefan** for organising the Ph.D. program and making it run smoothly.

I also would like to thank entire DBB, specially the GvH, JdG, RD, TH, EG and IN groups for creating an outstanding environment in our department.

Lastly, I would like to express my gratitude towards my **family** and **friends** who have supported me in all circumstances.

References

1. Vickery, H. B. The Origin of the Word Protein. *Yale J. Biol. Med.* **22,** 387–393 (1950).
2. Barroso, I. *et al.* Dominant negative mutations in human PPARγ associated with severe insulin resistance, diabetes mellitus and hypertension. *Nature* **402,** 880–883 (1999).
3. Chen, S., Sayana, P., Zhang, X. & Le, W. Genetics of amyotrophic lateral sclerosis: an update. *Mol. Neurodegener.* **8,** 28 (2013).
4. Mayosi, B. M. *et al.* Identification of Cadherin 2 (CDH2) Mutations in Arrhythmogenic Right Ventricular Cardiomyopathy. *Circ. Cardiovasc. Genet.* **10,** (2017).
5. Blackstone, E. A. & Joseph, P. F. The Economics of Biosimilars. *Am. Health Drug Benefits* **6,** 469–478 (2013).
6. Walsh, G. Biopharmaceutical benchmarks 2014. *Nat. Biotechnol.* **32,** 992–1000 (2014).
7. Corchero, J. L. *et al.* Unconventional microbial systems for the cost-efficient production of high-quality protein therapeutics. *Biotechnol. Adv.* **31,** 140–153 (2013).
8. Singh, R., Kumar, M., Mittal, A. & Mehta, P. K. Microbial enzymes: industrial progress in 21st century. *3 Biotech* **6,** 174 (2016).
9. Dunkle, J. A. *et al.* Structures of the bacterial ribosome in classical and hybrid states of tRNA binding. *Science* **332,** 981–984 (2011).
10. Schmeing, TM., Ramakrishnan, V. What recent ribosome structures have revealed about the mechanism of translation. *Nature* **461,** 1234–42 (2009).
11. Frank, J., Gao, H., Sengupta, J., Gao, N. & Taylor, D. J. The process of mRNA–tRNA translocation. *Proc. Natl. Acad. Sci.* **104,** 19671–19678 (2007).
12. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324,** 218–223 (2009).
13. Milón, P. & Rodnina, M. V. Kinetic control of translation initiation in bacteria. *Crit. Rev. Biochem. Mol. Biol.* **47,** 334–348 (2012).
14. Laursen, B. S., Sørensen, H. P., Mortensen, K. K. & Sperling-Petersen, H. U. Initiation of Protein Synthesis in Bacteria. *Microbiol. Mol. Biol. Rev.* **69,** 101–123 (2005).
15. Tomšic, J. *et al.* Late events of translation initiation in bacteria: a kinetic analysis. *EMBO J.* **19,** 2127–2136 (2000).
16. Gualerzi, C., Risuleo, G. & Pon, C. L. Initial rate kinetic analysis of the mechanism of initiation complex formation and the role of initiation factor IF-3. *Biochemistry (Mosc.)* **16,** 1684–1689 (1977).
17. Lukjancenko, O., Wassenaar, T. M. & Ussery, D. W. Comparison of 61 Sequenced Escherichia coli Genomes. *Microb. Ecol.* **60,** 708–720 (2010).

18. Reeve, B., Hargest, T., Gilbert, C. & Ellis, T. Predicting Translation Initiation Rates for Designing Synthetic Biology. *Front. Bioeng. Biotechnol.* **2,** (2014).
19. Espah Borujeni, A. *et al.* Precise quantification of translation inhibition by mRNA structures that overlap with the ribosomal footprint in N-terminal coding sequences. *Nucleic Acids Res.* **45,** 5437–5448 (2017).
20. Burkhardt, D. H. *et al.* Operon mRNAs are organized into ORF-centric structures that predict translation efficiency. *eLife* **6,**
21. Ding, Y. *et al.* In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature* **505,** 696–700 (2014).
22. Scharff, L. B., Childs, L., Walther, D. & Bock, R. Local absence of secondary structure permits translation of mRNAs that lack ribosome-binding sites. *PLoS Genet.* **7,** e1002155 (2011).
23. Bentele, K., Saffert, P., Rauscher, R., Ignatova, Z. & Blüthgen, N. Efficient translation initiation dictates codon usage at gene start. *Mol. Syst. Biol.* **9,** 675 (2013).
24. Dekel, E. & Alon, U. Optimality and evolutionary tuning of the expression level of a protein. *Nature* **436,** 588 (2005).
25. Shine, J. & Dalgarno, L. Determinant of cistron specificity in bacterial ribosomes. *Nature* **254,** 34–38 (1975).
26. Osterman, I. A., Evfratov, S. A., Sergiev, P. V. & Dontsova, O. A. Comparison of mRNA features affecting translation initiation and reinitiation. *Nucleic Acids Res.* **41,** 474–486 (2013).
27. Omotajo, D., Tate, T., Cho, H. & Choudhary, M. Distribution and diversity of ribosome binding sites in prokaryotic genomes. *BMC Genomics* **16,** (2015).
28. Shine, J. & Dalgarno, L. 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 (1974).
29. Salis, H. M., Mirsky, E. A. & Voigt, C. A. Automated Design of Synthetic Ribosome Binding Sites to Precisely Control Protein Expression. *Nat. Biotechnol.* **27,** 946–950 (2009).
30. Mutalik, V. K. *et al.* Precise and reliable gene expression via standard transcription and translation initiation elements. *Nat. Methods* **10,** 354–360 (2013).
31. Zelcbuch, L. *et al.* Spanning high-dimensional expression space using ribosome-binding site combinatorics. *Nucleic Acids Res.* **41,** e98 (2013).
32. Bonde, M. T. *et al.* Predictable tuning of protein expression in bacteria. *Nat. Methods* **13,** 233–236 (2016).
33. Chen, H., Bjerknes, M., Kumar, R. & Jay, E. Determination of the optimal aligned spacing between the Shine-Dalgarno sequence and the translation initiation codon of Escherichia coli mRNAs. *Nucleic Acids Res.* **22,** 4953–4957 (1994).
34. Liebeton, K., Lengefeld, J. & Eck, J. The nucleotide composition of the spacer sequence influences the expression yield of heterologously expressed genes in Bacillus subtilis. *J. Biotechnol.* **191,** 214–220 (2014).
35. Matteucci, M. D. & Heyneker, H. L. Targeted random mutagenesis: the use of ambiguously synthesized oligonucleotides to mutagenize sequences immediately 5′ of an ATG initiation codon. *Nucleic Acids Res.* **11,** 3113–3121 (1983).
36. Hui, A., Hayflick, J., Dinkelspiel, K. & de Boer, H. A. Mutagenesis of the three bases preceding the start codon of the beta-galactosidase mRNA and its effect on translation in Escherichia coli. *EMBO J.* **3,** 623–629 (1984).

37. Kudla, G., Murray, A. W., Tollervey, D. & Plotkin, J. B. Coding-sequence determinants of gene expression in Escherichia coli. *Science* **324,** 255–258 (2009).
38. Mortimer, S. A., Kidwell, M. A. & Doudna, J. A. Insights into RNA structure and function from genome-wide studies. *Nat. Rev. Genet.* **15,** 469–479 (2014).
39. Hecht, A. *et al.* Measurements of translation initiation from all 64 codons in E. coli. *Nucleic Acids Res.* **45,** 3615–3626 (2017).
40. Villegas, A. & Kropinski, A. M. An analysis of initiation codon utilization in the Domain Bacteria – concerns about the quality of bacterial genome annotation. *Microbiology* **154,** 2559–2661 (2008).
41. Chengguang, H. *et al.* Ribosomal selection of mRNAs with degenerate initiation triplets. *Nucleic Acids Res.* **45,** 7309–7325 (2017).
42. Looman, A. C. *et al.* Influence of the codon following the AUG initiation codon on the expression of a modified lacZ gene in Escherichia coli. *EMBO J.* **6,** 2489–2492 (1987).
43. Gonzalez de Valdivia, E. I. & Isaksson, L. A. A codon window in mRNA downstream of the initiation codon where NGG codons give strongly reduced gene expression in Escherichia coli. *Nucleic Acids Res.* **32,** 5198–5205 (2004).
44. Nørholm, M. H. H. *et al.* Improved production of membrane proteins in Escherichia coli by selective codon substitutions. *FEBS Lett.* **587,** 2352–2358 (2013).
45. Tuller, T. *et al.* An Evolutionarily Conserved Mechanism for Controlling the Efficiency of Protein Translation. *Cell* **141,** 344–354 (2010).
46. Goodman, D. B., Church, G. M. & Kosuri, S. Causes and effects of N-terminal codon bias in bacterial genes. *Science* **342,** 475–479 (2013).
47. Sengupta, J., Agrawal, R. K. & Frank, J. Visualization of protein S1 within the 30S ribosomal subunit and its interaction with messenger RNA. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 11991–11996 (2001).
48. Boni, I. V., Lsaeva, D. M., Musychenko, M. L. & Tzareva, N. V. Ribosome-messenger recognition: mRNA target sites for ribosomal protein S1. *Nucleic Acids Res.* **19,** 155–162 (1991).
49. Takahashi, S., Furusawa, H., Ueda, T. & Okahata, Y. Translation enhancer improves the ribosome liberation from translation initiation. *J. Am. Chem. Soc.* **135,** 13096–13106 (2013).
50. Vimberg, V., Tats, A., Remm, M. & Tenson, T. Translation initiation region sequence preferences in *Escherichia coli*. *BMC Mol. Biol.* **8,** 100 (2007).
51. Hirokawa, G., Demeshkina, N., Iwakura, N., Kaji, H. & Kaji, A. The ribosome-recycling step: consensus or controversy? *Trends Biochem. Sci.* **31,** 143–149 (2006).
52. Jewett, M. C., Miller, M. L., Chen, Y. & Swartz, J. R. Continued Protein Synthesis at Low [ATP] and [GTP] Enables Cell Adaptation during Energy Limitation. *J. Bacteriol.* **191,** 1083–1091 (2009).
53. Silhavy, T. J., Kahne, D. & Walker, S. The Bacterial Cell Envelope. *Cold Spring Harb. Perspect. Biol.* **2,** (2010).
54. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305,** 567–580 (2001).
55. Daley, D. O. *et al.* Global topology analysis of the *Escherichia coli* inner membrane proteome. *Science* **308,** 1321–1323 (2005).

56. Vollmer, W., Blanot, D. & de Pedro, M. A. Peptidoglycan structure and architecture. *FEMS Microbiol. Rev.* **32,** 149–167 (2008).
57. Weiner, J. H. & Li, L. Proteome of the *Escherichia coli* envelope and technological challenges in membrane proteome analysis. *Biochim. Biophys. Acta* **1778,** 1698–1713 (2008).
58. Georgopapadakou, N. H. & Liu, F. Y. Penicillin-binding proteins in bacteria. *Antimicrob. Agents Chemother.* **18,** 148–157 (1980).
59. Merdanovic, M., Clausen, T., Kaiser, M., Huber, R. & Ehrmann, M. Protein quality control in the bacterial periplasm. *Annu. Rev. Microbiol.* **65,** 149–168 (2011).
60. Plummer, A. M. & Fleming, K. G. From Chaperones to the Membrane with a BAM! *Trends Biochem. Sci.* **41,** 872–882 (2016).
61. Tsirigotaki, A., De Geyter, J., Šoštaric, N., Economou, A. & Karamanou, S. Protein export through the bacterial Sec pathway. *Nat. Rev. Microbiol.* **15,** 21–36 (2017).
62. Luirink, J., von Heijne, G., Houben, E. & de Gier, J.-W. Biogenesis of inner membrane proteins in *Escherichia coli*. *Annu. Rev. Microbiol.* **59,** 329–355 (2005).
63. Kudva, R. *et al.* Protein translocation across the inner membrane of Gram-negative bacteria: the Sec and Tat dependent protein transport pathways. *Res. Microbiol.* **164,** 505–534 (2013).
64. von Heijne, G. The signal peptide. *J. Membr. Biol.* **115,** 195–201 (1990).
65. Huber, D. *et al.* Use of thioredoxin as a reporter to identify a subset of Escherichia coli signal sequences that promote signal recognition particle-dependent translocation. *J. Bacteriol.* **187,** 2983–2991 (2005).
66. Steiner, D., Forrer, P., Stumpp, M. T. & Plückthun, A. Signal sequences directing cotranslational translocation expand the range of proteins amenable to phage display. *Nat. Biotechnol.* **24,** 823–831 (2006).
67. Lee, H. C. & Bernstein, H. D. The targeting pathway of *Escherichia coli* presecretory and integral membrane proteins is specified by the hydrophobicity of the targeting signal. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 3471–3476 (2001).
68. Estrozi, L. F., Boehringer, D., Shan, S., Ban, N. & Schaffitzel, C. Cryo-EM structure of the *E. coli* translating ribosome in complex with SRP and its receptor. *Nat. Struct. Mol. Biol.* **18,** 88–90 (2011).
69. Draycheva, A., Bornemann, T., Ryazanov, S., Lakomek, N.-A. & Wintermeyer, W. The bacterial SRP receptor, FtsY, is activated on binding to the translocon. *Mol. Microbiol.* **102,** 152–167 (2016).
70. Shan, S., Chandrasekar, S. & Walter, P. Conformational changes in the GTPase modules of the signal reception particle and its receptor drive initiation of protein translocation. *J. Cell Biol.* **178,** 611–620 (2007).
71. Bornemann, T., Holtkamp, W. & Wintermeyer, W. Interplay between trigger factor and other protein biogenesis factors on the ribosome. *Nat. Commun.* **5,** 4180 (2014).
72. Knoblauch, N. T. *et al.* Substrate specificity of the SecB chaperone. *J. Biol. Chem.* **274,** 34219–34225 (1999).
73. Mitra, K., Frank, J. & Driessen, A. Co- and post-translational translocation through the protein-conducting channel: analogous mechanisms at work? *Nat. Struct. Mol. Biol.* **13,** 957–964 (2006).
74. Tani, K., Shiozuka, K., Tokuda, H. & Mizushima, S. In vitro analysis of the process of translocation of OmpA across the *Escherichia coli* cytoplasmic membrane. A trans-

location intermediate accumulates transiently in the absence of the proton motive force. *J. Biol. Chem.* **264,** 18582–18588 (1989).
75. Huber, D. *et al.* SecA interacts with ribosomes in order to facilitate posttranslational translocation in bacteria. *Mol. Cell* **41,** 343–353 (2011).
76. Brundage, L., Hendrick, J. P., Schiebel, E., Driessen, A. J. & Wickner, W. The purified E. coli integral membrane protein SecY/E is sufficient for reconstitution of SecA-dependent precursor protein translocation. *Cell* **62,** 649–657 (1990).
77. Komar, A. A. A pause for thought along the co-translational folding pathway. *Trends Biochem. Sci.* **34,** 16–24 (2009).
78. Ciryam, P., Morimoto, R. I., Vendruscolo, M., Dobson, C. M. & O'Brien, E. P. In vivo translation rates can substantially delay the cotranslational folding of the Escherichia coli cytosolic proteome. *Proc. Natl. Acad. Sci.* **110,** E132–E140 (2013).
79. Nilsson, O. B. *et al.* Cotranslational Protein Folding inside the Ribosome Exit Tunnel. *Cell Rep.* **12,** 1533–1540 (2015).
80. Kim, S. J. *et al.* Protein folding. Translational tuning optimizes nascent protein folding in cells. *Science* **348,** 444–448 (2015).
81. Ugrinov, K. G. & Clark, P. L. Cotranslational Folding Increases GFP Folding Yield. *Biophys. J.* **98,** 1312–1320 (2010).
82. Evans, M. S., Sander, I. M. & Clark, P. L. Cotranslational folding promotes beta-helix formation and avoids aggregation in vivo. *J. Mol. Biol.* **383,** 683–692 (2008).
83. Zhang, G., Hubalewska, M. & Ignatova, Z. Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nat. Struct. Mol. Biol.* **16,** 274–280 (2009).
84. Hoffmann, A., Bukau, B. & Kramer, G. Structure and function of the molecular chaperone Trigger Factor. *Biochim. Biophys. Acta* **1803,** 650–661 (2010).
85. Kim, Y. E., Hipp, M. S., Bracher, A., Hayer-Hartl, M. & Hartl, F. U. Molecular chaperone functions in protein folding and proteostasis. *Annu. Rev. Biochem.* **82,** 323–355 (2013).
86. Hayer-Hartl, M., Bracher, A. & Hartl, F. U. The GroEL-GroES Chaperonin Machine: A Nano-Cage for Protein Folding. *Trends Biochem. Sci.* **41,** 62–76 (2016).
87. Mattanovich, D. *et al.* Recombinant protein production in yeasts. *Methods Mol. Biol. Clifton NJ* **824,** 329–358 (2012).
88. van Dijl, J. M. & Hecker, M. *Bacillus subtilis*: from soil bacterium to super-secreting cell factory. *Microb. Cell Factories* **12,** 3 (2013).
89. Rosano, G. L. & Ceccarelli, E. A. Recombinant protein expression in *Escherichia coli*: advances and challenges. *Front. Microbiol.* **5,** 172 (2014).
90. Johnson, I. S. Human insulin from recombinant DNA technology. *Science* **219,** 632–637 (1983).
91. Ferrer-Miralles, N., Domingo-Espín, J., Corchero, J. L., Vázquez, E. & Villaverde, A. Microbial factories for recombinant pharmaceuticals. *Microb. Cell Factories* **8,** 17 (2009).
92. Recombinant protein production in bacterial hosts. *Drug Discov. Today* **19,** 590–601 (2014).
93. Bentley, W. E., Mirjalili, N., Andersen, D. C., Davis, R. H. & Kompala, D. S. Plasmid-encoded protein: the principal factor in the 'metabolic burden' associated with recombinant bacteria. *Biotechnol. Bioeng.* **35,** 668–681 (1990).

94. Soriano, E., Borth, N., Katinger, H. & Mattanovich, D. Flow cytometric analysis of metabolic stress effects due to recombinant plasmids and proteins in *Escherichia coli* production strains. *Metab. Eng.* **1,** 270–274 (1999).
95. Kudla, G., Lipinski, L., Caffin, F., Helwak, A. & Zylicz, M. High Guanine and Cytosine Content Increases mRNA Levels in Mammalian Cells. *PLoS Biol.* **4,** (2006).
96. Thomas, J. G. & Baneyx, F. Protein misfolding and inclusion body formation in recombinant *Escherichia coli* cells overexpressing Heat-shock proteins. *J. Biol. Chem.* **271,** 11141–11147 (1996).
97. Kane, J. F. Effects of rare codon clusters on high-level expression of heterologous proteins in *Escherichia coli*. *Curr. Opin. Biotechnol.* **6,** 494–500 (1995).
98. Mujacic, M., Cooper, K. W. & Baneyx, F. Cold-inducible cloning vectors for low-temperature protein expression in *Escherichia coli*: application to the production of a toxic and proteolytically sensitive fusion protein. *Gene* **238,** 325–332 (1999).
99. Gopal, G. J. & Kumar, A. Strategies for the production of recombinant protein in *Escherichia coli*. *Protein J.* **32,** 419–425 (2013).
100. Chamberlin, M., McGrath, J. & Waskell, L. New RNA polymerase from *Escherichia coli* infected with bacteriophage T7. *Nature* **228,** 227–231 (1970).
101. Schlegel, S. *et al.* Optimizing membrane protein overexpression in the *Escherichia coli* strain Lemo21(DE3). *J. Mol. Biol.* **423,** 648–659 (2012).
102. Yang, Z. *et al.* Highly efficient production of soluble proteins from insoluble inclusion bodies by a two-step-denaturing and refolding method. *PloS One* **6,** e22981 (2011).
103. Wacker, M. *et al.* N-linked glycosylation in *Campylobacter jejuni* and its functional transfer into *E. coli*. *Science* **298,** 1790–1793 (2002).
104. Chen, X. *et al.* Metabolic engineering of *Escherichia coli*: a sustainable industrial platform for bio-based chemical production. *Biotechnol. Adv.* **31,** 1200–1223 (2013).
105. Sørensen, H. P. & Mortensen, K. K. Advanced genetic strategies for recombinant protein expression in *Escherichia coli*. *J. Biotechnol.* **115,** 113–128 (2005).
106. Gustafsson, C. *et al.* Engineering Genes for Predictable Protein Expression. *Protein Expr. Purif.* **83,** 37–46 (2012).
107. Wallin, E. & von Heijne, G. Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci. Publ. Protein Soc.* **7,** 1029–1038 (1998).
108. Lundstrom, K. Structural genomics and drug discovery. *J. Cell. Mol. Med.* **11,** 224–238 (2007).
109. Wagner, S. *et al.* Consequences of membrane protein overexpression in *Escherichia coli*. *Mol. Cell. Proteomics MCP* **6,** 1527–1550 (2007).
110. Xu, L. Y. & Link, A. J. Stress responses to heterologous membrane protein expression in *Escherichia coli*. *Biotechnol. Lett.* **31,** 1775–1782 (2009).
111. Glick, B. R. Metabolic load and heterologous gene expression. *Biotechnol. Adv.* **13,** 247–261 (1995).
112. Studer, S. M. & Joseph, S. Unfolding of mRNA secondary structure by the bacterial translation initiation complex. *Mol. Cell* **22,** 105–115 (2006).
113. Kozak, M. Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene* **361,** 13–37 (2005).

114. Seo, S. W. *et al.* Predictive design of mRNA translation initiation region to control prokaryotic translation efficiency. *Metab. Eng.* **15,** 67–74 (2013).
115. Na, D., Lee, S. & Lee, D. Mathematical modeling of translation initiation for the estimation of its efficiency to computationally design mRNA sequences with desired expression levels in prokaryotes. *BMC Syst. Biol.* **4,** 71 (2010).
116. Hjelm, A. Optimizing membrane and secretory protein production in Gram-negative bacteria. *DIVA* (2015).
117. Raetz, C. R. H. & Whitfield, C. Lipopolysaccharide Endotoxins. *Annu. Rev. Biochem.* **71,** 635–700 (2002).