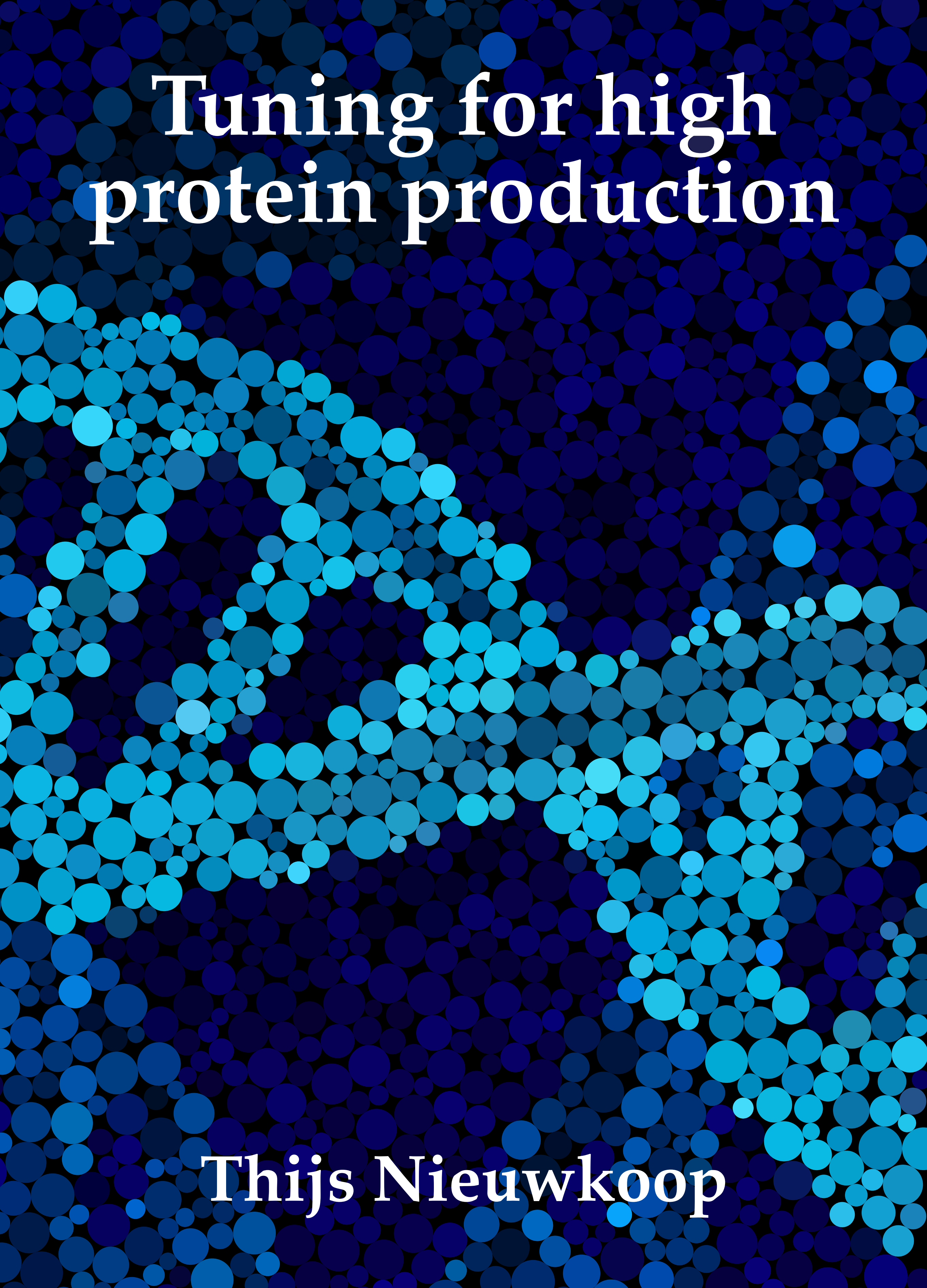
Tuning for high protein production



Thijs Nieuwkoop

Propositions

1. mRNA secondary structures are changeable without altering the codon usage but not vice versa.
   (this thesis)

2. N-terminal purification tags unintendedly standardized gene expression rates.
   (this thesis)

3. The scientific community needs an agreed-upon redefinition for the word ″gene″.

4. The success rate of codon optimization algorithms suffers from positive publication bias.

5. Programming skills are becoming essential in science due to the rise of big data and should be included in the curriculum.

6. Natural selection in humans has drastically changed and no longer selects for the ″fittest″.

Propositions belonging to the thesis entitled:

″Tuning for high protein production″

Thijs Nieuwkoop
Wageningen, 8th December 2021

Tuning for high protein production

Thijs Nieuwkoop

Thesis committee

Promotor
Prof. Dr John van der Oost
Professor of Microbial Genetics
Wageningen University & Research

Co-promoter
Dr Nico J.P.H. Claassens
Assistant Professor, Laboratory of Microbiology
Wageningen University & Research

Other members
Prof. Dr Dick de Ridder, Wageningen University & Research
Dr Jan-Willem de Gier, Stockholm University, Sweden
Dr Jan Schouten, Byondis B.V., Nijmegen
Dr Markus Jeschek, ETH Zürich, Switzerland

This research was conducted under the auspices of the Graduate School VLAG (Advanced studies in Food Technology, Agrobiotechnology, Nutrition and Health Sciences).

Tuning for high protein production

Thijs Nieuwkoop

Thesis
submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of Rector Magnificus,
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Wednesday 8 December 2021
at 4 p.m. in the Aula.

Thijs Nieuwkoop
Tuning for high protein production
186 pages

PhD thesis, Wageningen University, Wageningen, The Netherlands (2021)
With references, with summary in English

ISBN: 978-94-6395-999-5
DOI: https://doi.org/10.18174/554205

Table of contents

List of Abbreviations

3′UTR	**Three Prime** **U**n**t**ranslated **R**egion
5′UTR	**Five Prime** **U**n**t**ranslated **R**egion
AU	**A**rbitrary **U**nit
BCD	**B**i**c**istronic **D**esign
BPP	**B**ase **P**airing **P**robability
CAI	**C**odon **A**daptation **I**ndex
CDS	**C**o**d**ing **S**equence
CSC	**C**odon **S**tability **C**oefficient
FACS	**F**luorescence-**A**ctivated **C**ell **S**orting
FM	**F**unctional **M**utant
FS	**F**rame **S**hift
GFP	**G**reen **F**luorescent **P**rotein
H	**H**armonized
IGR	**I**nter**g**enic **R**egion
IRES	**I**nternal **R**ibosome **E**ntry **S**ite
LASSO	**L**east **A**bsolute **S**hrinkage and **S**election **O**perator
MCD	**M**ono**c**istronic **D**esign
mRNA	**M**essenger **RNA**
ORF	**O**pen **R**eading **F**rame
PABP	**P**oly(**A**)-**B**inding **P**rotein
POI	**P**rotein **O**f **I**nterest
PSAT	*Post* **S**top *Ante* **T**erminator
RBS	**R**ibosome **B**inding **S**ite
RFP	**R**ed **F**luorescent **P**rotein
RFR	**R**andom **F**orest **R**egressor
RNAP	**RNA** **P**olymerase
RU	**R**egularly **U**sed
SD	**S**hine-**D**algarno
WT	**W**ild **T**ype

Chapter 1

General introduction and Thesis outline

1.1 Biotechnology and protein production

Nearly all the catalytic activity in a biological cell can be attributed back to proteins. These enzymes are marvellous nanoscale machines that have obtained astounding functions over the course of evolution. Pathways and networks of proteins catalyze the metabolic conversion of substrates to products and thereby release the substrate's inherent stored energy while generating building blocks that can be used by other proteins to catalyze the biosynthesis of new products (Alberts et al., 2009). These metabolic conversions form the basis of early biotechnological applications, mainly fermentation to produce alcoholic drinks and dairy food, which have been documented dating back to 7000 BCE (McGovern et al., 2004).

Only in the 19th-century scientists became aware that proteins are responsible for the observed catalysis. They also discovered that some proteins function as signalling molecules (hormones) within living organisms and can be used as pharmaceuticals. An early bio-based pharmaceutical was porcine insulin, which was extracted from pig pancreases. Two hundred grams of porcine insulin required about 2000 kilograms of pig organs (Wendt, 2013). The inefficiency of harvesting natural protein sources as well as the advances in genetic engineering prompted the more efficient production of insulin in engineered microbial cells. Compared to natural sources, a much higher product to cell biomass ratio is generally obtained by engineering a microorganism to produce a specific protein. Human insulin was the first heterologous produced protein (Goeddel et al., 1979). The gene encoding human insulin was introduced into *Escherichia coli*, a common bacterium in the gut of mammals and a very popular model bacterium.

Defining a gene. *The definition of a "gene" has become blurred since it was first coined in 1909 by Wilhelm Johannsen (Johannsen, 1909). Originally a "gene" was formulated as a "unit of heredity". The concept changed around the 1960s to "a continuous segment of DNA sequence specifying a polypeptide chain". However, we now know that a single gene can encode multiple mRNAs via, e.g., alternative splicing and that regulatory elements do not need to be physically contiguous. There is still no generally agreed-upon definition but for a more in-depth analysis on the history of the usage of the term "gene" and a modern definition, I refer to a perspective article by Portin and Wilkins (Portin and Wilkins, 2017). For all intents and purposes, I define a gene here as the entire DNA sequence required for the synthesis of a specific protein or RNA, including the transcriptional and translational regulatory elements required for proper biological functioning of the protein or RNA.*

Optimized heterologous protein production in microorganisms can reach yields of more than 50% of the total protein content in the cell (Mierendorf et al., 1998). This production efficiency indicates the potential of heterologous protein production as an alternative for homologous production by natural organisms. Obtaining a consistently high yield of heterologous protein production is a major problem that has still not been solved. Due to the many interconnected molecular features that all influence the overall protein yield, it is most likely that optimization of one of these features results in negatively affecting others. Study of the regulatory elements within genes has identified fundamental mechanisms that regulate transcription and translation rates. However, in practice, the tuning of these regulatory elements is highly unpredictable.

In this thesis, I describe the exploration of regulatory genetic features that contribute to overall protein production to better understand and predict optimal variants. Additionally, genetic modules are explored to minimize known limiting factors in protein production to simultaneously offer standardized genetic parts and aid in the disentanglement of the interconnected features for fundamental studies. Despite the apparent fundamental nature of this project, it is also highly applicable. Results can be directly applied to current and future protein production pipelines to improve production efficiencies or make future production plans viable.

1.2 From DNA to protein

The flow of information within biological cells goes from nucleic acids to protein. The central dogma, formulated by Francis Crick in 1958, describes the hypothesis that the flow of information is only possible in one way from nucleic acids to protein and that once the information ends up as a protein, it is not possible to be converted back to nucleic acids (Crick, 1958). James Watson later expanded this hypothesis by stating that there are two irreversible information flows, namely information on DNA would only flow to RNA and from RNA to protein and both flows could not be reversed (Watson, 1965). However, the latter formulation is now proven incorrect. Many viruses are capable of reverse transcription, where they synthesize DNA from RNA or have RNA to RNA replication systems (Figure 1.1, asterisk) (Menéndez-Arias, Sebastián-Martín, and Álvarez, 2017). Crick's central dogma still proves true to this day as no examples are found where the "information" contained in a protein flows back into RNA or DNA.

A cell is a non-fixed system that goes through several distinct phases during the cell cycle and can react to environmental changes. To adjust and respond to internal and external changes, a cell needs to alter its internal protein composition to execute different processes. Proteins are continuously produced and broken down within the cell. A specific turnover rate is achieved by regulating the amount of production and degradation of a particular protein. In a human cell, protein half-lives range from 45 minutes to 22.5 hours (Eden et al., 2011). This kind of modulation is required for a well-functioning, healthy cell. The cells of the human body continuously regulate the expression levels of about 21 thousand genes. In order to become a specific cell type and carry out particular functions, variable sub-sets of the total proteome are produced in each cell (Salzberg, 2018).

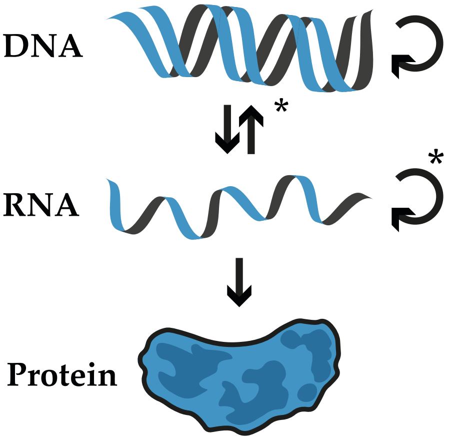


FIGURE 1.1: **The Central Dogma.** The flow of genetic information goes primarily from DNA to RNA to Protein. DNA to DNA replication occurs in all cellular life forms, as well as in DNA viruses, whereas RNA to RNA replication and RNA to DNA reverse transcription (indicated by an asterisk) occurs in RNA viruses.

These native modulations can be manipulated for heterologous protein production to trick the cell into producing high amounts of a specific protein. In a cell, there are three distinct levels of regulation. The transcriptional process dictates the amount of RNA that is transcribed from DNA at a specific time, whereas the translational process controls the amount of protein that will be produced from a messenger RNA (mRNA) molecule. Translation can be split into two distinct phases: translation initiation and translation elongation. During translation initiation, the translational machinery assembles on the mRNA. During translation elongation, the translational machinery travels over the mRNA, translating the nucleotide information to the amino acid sequence, thereby synthesizing the protein. Both phases have a multitude of factors that can enhance or limit the overall efficiency of the translational process. Finally, the protein turnover is dictated by the protein's inherent stability and the amount of protease activity within the cell.

In order to achieve high heterologous protein production, understanding and consequent optimization and tuning of all three levels are required (Figure 1.2). Transcription is mainly regulated by the binding of RNA polymerase. This polymerase complex binds to a specific sequence called the promoter sequence. Weaker promoter sequences consequently lead to less efficient binding of the transcriptional machinery. Through these different promoter sequences, the base level of transcription is modulated. Eukaryotes possess an additional, more nuanced level of control over the rate of transcription by compacting large parts of their genome into tight higher-order structures (heterochromatin), thereby preventing transcription via spatial restrictions. Unlike prokaryotes, the transcription of genes in eukaryotes is normally in the off-state. Upon unwinding of these higher-order structures by histone remodelling enzymes, genes locally become available for transcription. Prokaryotic genes are generally in a transcriptional on-state and only by the involvement of repressors and activators additional control over the transcription is exerted besides the promoter strength. The transcription stops when the RNA polymerase encounters a termination signal. This termination process is important both to prevent the unintended transcription of downstream genes and for the stabilization of the mRNA.

In prokaryotes, termination of the transcription process can occur in two ways: Rho-dependent and Rho-independent. Rho is a helicase protein that binds to cytosine-rich RNA and induces a termination process. Rho-independent termination does not depend on any protein. A strong secondary stem structure of the RNA transcript is formed during the transcription of these intrinsic terminators, followed by a uracil-rich stretch. This structure halts the RNA polymerase and causes dissociation of the complex. In eukaryotes, multiple RNA polymerases exist. Surprisingly, the primary RNA polymerase II does not recognize a specific transcription stop site and seemingly halts at a random position past the transcribed gene. However, a specific polyA-signal (e.g. AAUAAA) is recognized in the 3′ untranslated region (3′UTR) on the synthesized RNA stretch by a polyadenylation complex. This complex cleaves the mRNA to its intended length and adds a stretch of adenines to the 3′UTR of the mRNA. Both the adenine-tail and rho-independent terminator structure have an additional important function besides the transcription termination, in that both stabilize the mRNA against RNase degradation and thus influence the amount of protein that can be translated from a single mRNA.

The promoter strength, genome location and terminator can all be used to influence the transcriptional efficiency in heterologous protein production. However, if the goal is high and efficient protein production, striving for the highest possible transcription can be counterproductive. If the translational efficiency is

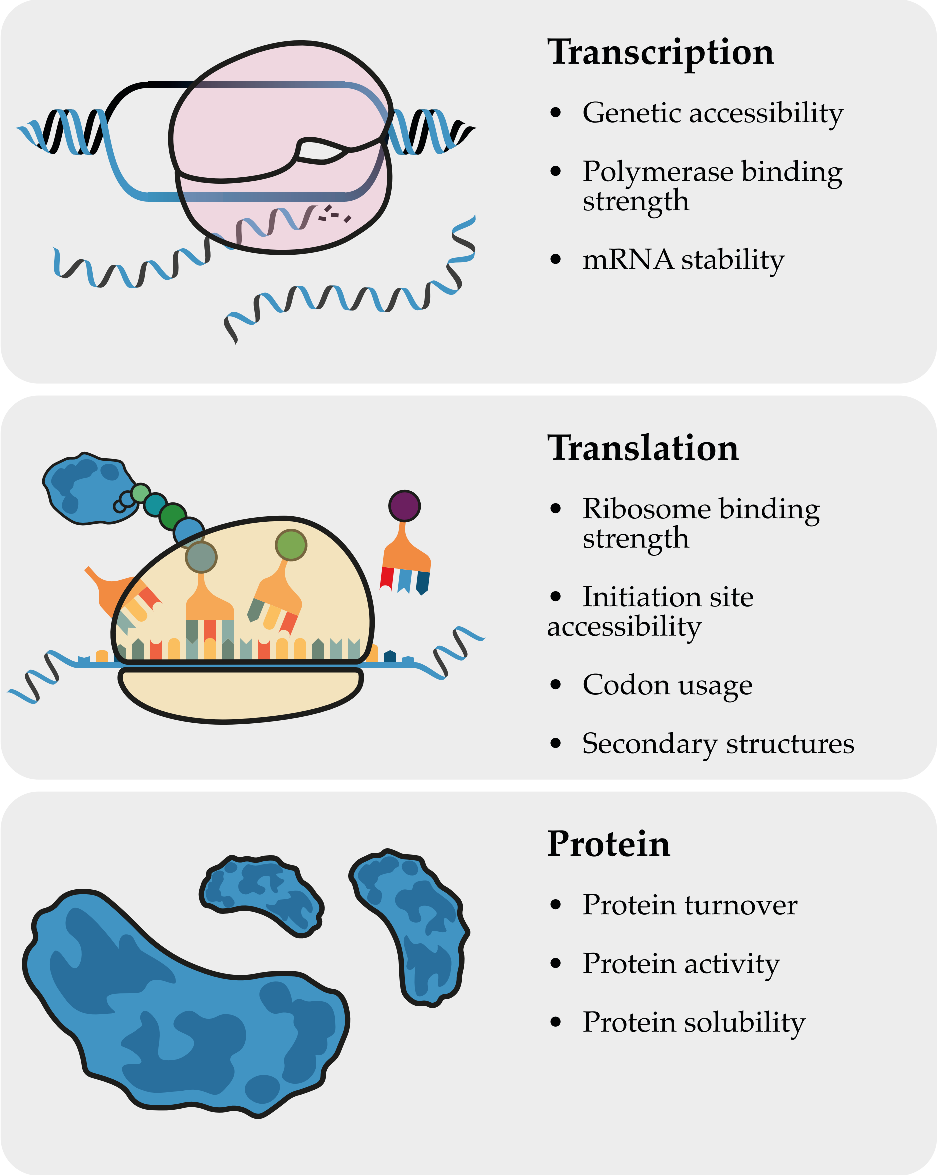


FIGURE 1.2: **The main regulatory features affecting gene expression at the transcription, translation and protein phase.**

limiting, increasing the levels of mRNA will not contribute to increasing production. Instead, energy loss will occur due to the synthesis of unused mRNA. Based on the translational capability, an optimal transcriptional level will exist.

After transcription, an mRNA can be translated by a ribosome into protein. The translation initiation process is fundamentally different between prokaryotes and eukaryotes. Prokaryotes use a relatively simple mechanism in which the 16S ribosomal subunit recognizes the 5′ untranslated region (5′UTR) of an mRNA molecule, more specifically the Ribosome Binding Site (RBS) with its Shine-Dalgarno (SD) sequence, a 5-6 nucleotide organism-specific sequence located just upstream of the coding sequence of the gene. Apart from several initiation factor proteins that form a complex with the 30S ribosomal subunit, two main factors influence the rate of translation initiation in prokaryotes: the SD sequence's complementarity with the 3′ end of the organism's 16S rRNA, as well as the accessibility of the transcript's RBS/SD. If the SD sequence is involved in a secondary structure of the mRNA, the ribosome binds less efficiently. In practice, this is often the limiting factor in heterologous gene expression. Eukaryotic translation initiation is more complex as it involves initiation factor proteins as well as a cap-binding protein. This complex assembles on the 7-methylguanosine cap at the 5′UTR of mature eukaryotic mRNAs. Subsequently, the complex scans for the start of the open reading frame, aided by the presence of a Kozak sequence surrounding the start codon. Similarly as in prokaryotes, regulation of protein production by altering the 5′UTR appears to be an effective strategy to tune and improve eukaryotic protein production as translation initiation was shown to, generally, be the rate-limiting step in yeast (Shah et al., 2013).

In nature, there are 20 amino acids, which are the building blocks of protein. The order and identity of amino acids dictate a protein's shape and function. DNA consists of only four nucleotides: cytosine (C), guanine (G), adenine (A) and thymine (T). To identify how the nucleotides code for amino acids Nirenberg and Matthaei demonstrated back in 1961 that a stretch of uracil (non-methylated thymine used on the RNA level) resulted in a string of the amino acid phenylalanine (Nirenberg and Matthaei, 1961). Around the same time, Crick demonstrated that nucleotide triplets are the units (termed codons) that are translated one by one by cognate, charged tRNAs (Crick, 1988). By testing all combinations of three nucleic acids, the genetic code was eventually cracked (Nirenberg et al., 1966; Söll et al., 1966). Interestingly there are a total of 64 codons (4^3), 61 of which encode the 20 amino acids indicating there is a redundancy in the genetic code (Table 1.1). Most amino acids are encoded by multiple synonymous codons and these codons are not distributed equally throughout the genome. This unequal distribution of codons within an organism is called the codon usage bias. Codons are

TABLE 1.1: Codon table showing codon frequencies and tRNA gene copy numbers for *E.coli* (*Eco*), *Saccharomyces cerevisiae* (*Sce*) and *Homo sapiens* (*Hsa*). CDS data was obtained from RefSeq (015291845.1 (*Eco*), 000146045.2 (*Sce*), 000001405.39 (*Hsa*)). tRNA copy numbers were obtained from the GtRNAdb (http://gtrnadb.ucsc.edu). The first, second and third base of a codon point to the associated amino acid.

	Second base																
First base	**T**				**C**				**A**				**G**				**Third base**
	Amino acid	*Eco* Codon % (tRNA #)	*Sce* Codon % (tRNA #)	*Hsa* Codon % (tRNA #)	Amino acid	*Eco* Codon % (tRNA #)	*Sce* Codon % (tRNA #)	*Hsa* Codon % (tRNA #)	Amino acid	*Eco* Codon % (tRNA #)	*Sce* Codon % (tRNA #)	*Hsa* Codon % (tRNA #)	Amino acid	*Eco* Codon % (tRNA #)	*Sce* Codon % (tRNA #)	*Hsa* Codon % (tRNA #)	
T	Phe	57 (0)	60 (0)	57 (0)	Ser	15 (0)	26 (11)	15 (9)	Tyr	57 (0)	57 (0)	43 (0)	Cys	44 (0)	62 (0)	44 (0)	**T**
		43 (2)	40 (10)	43 (10)		15 (2)	16 (0)	28 (0)		43 (3)	43 (8)	57 (13)		56 (1)	38 (4)	56 (29)	**C**
	Leu	13 (1)	28 (7)	4 (4)		12 (1)	21 (3)	12 (4)	Stop	64 (0)	47 (0)	7 (0)	Stop	29 (0)	31 (0)	64 (0)	**A**
		13 (1)	28 (10)	13 (6)		15 (1)	10 (1)	15 (4)		7 (0)	22 (0)	29 (0)	Trp	100 (1)	100 (6)	100 (7)	**G**
C		10 (0)	13 (0)	10 (9)	Pro	16 (0)	31 (2)	53 (9)	His	57 (0)	64 (0)	57 (0)	Arg	38 (4)	14 (6)	10 (7)	**T**
		10 (1)	6 (1)	50 (0)		12 (1)	16 (0)	12 (0)		43 (1)	36 (7)	43 (9)		40 (0)	6 (0)	40 (0)	**C**
		4 (1)	14 (3)	10 (3)		19 (1)	41 (10)	16 (7)	Gln	35 (2)	69 (9)	35 (6)		6 (0)	7 (0)	2 (6)	**A**
		50 (4)	11 (0)	13 (9)		53 (1)	12 (0)	19 (4)		65 (2)	31 (1)	65 (13)		10 (1)	4 (1)	4 (4)	**G**
A	Ile	51 (0)	46 (13)	51 (15)	Thr	17 (0)	34 (11)	44 (9)	Asn	45 (0)	60 (0)	55 (0)	Ser	15 (0)	16 (0)	15 (0)	**T**
		42 (3)	26 (0)	42 (3)		44 (2)	21 (0)	27 (0)		55 (4)	40 (10)	45 (25)		28 (1)	11 (4)	15 (8)	**C**
		7 (0)	28 (2)	7 (5)		13 (1)	31 (4)	17 (6)	Lys	77 (6)	59 (7)	23 (12)	Arg	4 (1)	47 (11)	6 (6)	**A**
	Met Start	100 (6)	100 (10)	100 (20)		27 (2)	14 (1)	13 (5)		23 (0)	41 (14)	77 (15)		2 (1)	21 (1)	38 (5)	**G**
G	Val	26 (0)	39 (14)	15 (9)	Ala	16 (0)	37 (11)	21 (26)	Asp	63 (0)	65 (0)	63 (0)	Gly	34 (0)	46 (0)	34 (0)	**T**
		22 (2)	20 (0)	22 (0)		27 (2)	22 (0)	27 (0)		37 (3)	35 (16)	37 (13)		40 (4)	20 (16)	15 (14)	**C**
		15 (5)	22 (2)	37 (5)		21 (3)	30 (5)	36 (8)	Glu	69 (4)	70 (14)	31 (8)		11 (1)	23 (3)	40 (9)	**A**
		37 (0)	19 (2)	26 (13)		36 (0)	11 (0)	16 (4)		31 (0)	30 (2)	69 (8)		15 (1)	12 (2)	11 (5)	**G**

translated by the ribosome into amino acids. The ribosome facilitates the subsequent binding of tRNAs with anti-codons to the complementary codons on the mRNA sequence. tRNAs are molecules that are specifically charged with a single amino acid and recognize a specific codon. Interestingly, organisms do not have a full set of tRNAs to cover all possible codons (Table 1.1). Hence, some codons can only be translated by nonperfect complementary tRNAs via wobble base pair interactions or via modified tRNAs. Additionally, the intracellular concentration of each tRNA may vary during the cell cycle, in different cell types and/or as a response to external stimuli (Torrent et al., 2018). These changes in tRNA concentrations contribute to the proper cellular response by improving or decreasing the translation efficiency of individual genes.

A heterologous gene that should be highly expressed ideally uses codons that can be efficiently translated by the cell's tRNA pool. Besides the codon usage, there is another major factor that influences the efficiency of translation elongation. Due to hydrogen bonds between nucleotides, an mRNA generally folds itself into a thermodynamically stable conformation, e.g. simple stem-loop structures, or complicated variants thereof. These secondary and tertiary structures stabilize the mRNA and protect it against degradation. However, they can also have adverse effects on translation initiation and elongation. Structures involving the 5'UTR can block the ribosome from binding to the mRNA. Structures within the ORF can slow down or block the elongating ribosome as it needs to unfold these structures during translation. Optimizing secondary structures for heterologous protein production proves very difficult, especially because *in silico* tools are limited in their predictive power. Additionally, optimization of one factor, whether it is the secondary structure or codon usage, can have unintentional side effects. The codon usage is directly linked to secondary structure formations and vice versa. This makes gene optimization for high expression a very complex issue and explains why 60 years after the discovery of the genetic code, we are still not able to design the best-optimized gene for protein production.

1.3 Synthetic DNA synthesis and lab evolution

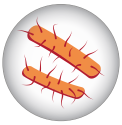
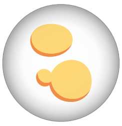
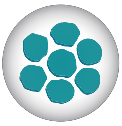
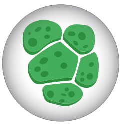
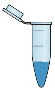
Advances in synthetic DNA synthesis and genetic engineering tools have shifted the codon research field towards massive high throughput experiments to generate big data. The cost of synthetic DNA synthesis has decreased consistently over the past 30 years from about 10 dollars per base to 5 cents per base (Hughes and Ellington, 2017)). This now allows for the generation of huge libraries of genetic variants to study the effects on protein yield. Before these advances, trends in, for example, codon usage were mainly obtained by studying proteomics and genomics data of native genes. However, while moderate trends were found, this data is not ideal to discover fundamental rules as different underlying features exist between different genes. With the current advances, genetic libraries have been generated of up to 100 million gene variants coding for the exact same protein (Boer et al., 2020). These approaches reveal fundamental expression rules, but they might also be limited to the protein in question.

Besides fundamental understanding of regulatory elements, the selection of high producing gene variants is also of interest. With an *in vivo* selection system, the best performing genetic construct could be automatically selected from a huge pool of possibilities. To achieve this kind of automatic selection, increased production of the protein of interest should give the cell some advantage over the population. Generally, increasing (heterologous) protein production results in negative cell fitness, as the cell needs to direct energy to production instead of growth. However, the protein production level can be tied to growth via genetic constructs or external selection pressure. By coupling the translation efficiency of a protein of interest to a growth advantage, indirect selection for production can occur, for example, via the production of an antibiotic resistance marker or an essential amino acid. However, a mutation inside the genetic coupling construct can result in false positives. A mutation that would disable the coupling effect and cause the production of the selection marker to be independent will be highly favourable. The cell no longer needs to produce the protein of interest, saving energy and gaining an additional growth advantage. An alternative to this selection pressure would be to couple the production of the protein of interest to the production of a seemingly non-essential gene such as a fluorescent protein. Instead of constant pressure during growth, selection can take place in the stationary phase using FACS. By sorting the cells with high fluorescence, selection occurs for high expression. While mutated cells that only produce the fluorescent marker will be sorted as high producers during the selection step, they do not have a growth advantage during the culturing phase and therefore are not enriched in the culture. With the rise of large-scale synthetic DNA synthesis and screening capabilities and the development of novel laboratory selection systems, an alternative has become available for the *a priori* design approach. Especially from an industrial standpoint, these kinds of approaches are attractive. Ideally, the screening/selection is directly made in the target host strain under production conditions to fully optimize the production process.

1.4 The protein market

The protein production market is steadily growing as new technologies are developed. Improvements in DNA modification, protein engineering, production yield and downstream purification processes all contribute to a market where we are no longer limited to natural proteins but are able to design and produce tailor-made proteins. The protein market can be split into three main categories: therapeutic proteins, industrial proteins and proteins for research purposes. The therapeutic protein market has the highest global value in which monoclonal antibodies occupy the largest share, followed by insulin (Sumant and Shaikh, 2017). Monoclonal antibodies are mainly used to treat cancer and autoimmune diseases. Insulin is a hormone used to treat type 1 diabetes. Vaccines are traditionally also based on protein production as they rely on the production of viral capsid protein or inactivated or attenuated viruses. Industrial proteins are mainly used in food production, animal feed and technical applications such as detergents and the paper industry. Meat replacers and alternatives are gaining attention as they offer a reduction in CO_2 emission and animal cruelty. Examples are the production of haemoglobin and myoglobin to simulate the meat taste in vegetarian burgers and the production of casein in yeast for vegan milk. Finally, a significant part of the heterologous proteins is used for research purposes such as diagnostics. Based on the currently observed trends, the protein market will grow in the coming years as new demands arise and innovations in production processes lower costs (Mordor Intelligence, 2020). Many different protein production platforms exist with each their advantages and disadvantages (Table 1.2). Prokaryotes and Yeast generally have high productivity while higher eukaryotic cells, like CHO or Human cell lines, are used if post-translational modifications are required. A major goal in the protein production field is to obtain a similar post-translation modification in prokaryotes and introduce efficient secretion systems to benefit both from the easy growth, high yields and ease of purification. Additionally, new techniques are constantly being developed to increase protein yields, such as the discovery or optimization of genetic parts, novel codon optimization approaches and improved culture conditions.

TABLE 1.2: Applications of common expression systems with general pros and cons. However, as many of these systems are actively improved there are strains that solve some of the cons mentioned. (Demain and Vaishnav, 2009, Based on a table presented in the Protein Expression Handbook by Thermo Fisher Scientific).

Expression system	**Common applications**	**Common Pros**	**Common Cons**
Bacterial 	• Fundamental protein studies • Expression of simple proteins • Antibody production	• Scalable • Rapid growth • Ease of culture • High protein productivity • Secretion capability (gram-positive)	• Inclusion body formation • No glycosylation • Endotoxins • Limited secretion capability (gram-negative)
Yeast 	• Fundamental protein studies • Expression of simple proteins • Antibody production	• Eukaryotic protein processing • Moderate secretion capability • High protein production • Ease of culture	• Hyper-glycosylation of N-linked sites • Fermentation required for high yields • Cell lysis more difficult
Insect 	• Fundamental protein studies • Expression of intracellular proteins • Expression of protein complexes • Virus production	• Similar to eukaryotic protein processing • Usable in static and suspension culture	• Baculovirus vector production more time consuming • More complex culture conditions than prokaryotes
Mammalian 	• Fundamental protein studies • Antibody production • Expression of protein complexes • Virus production	• Ideal for therapeutic proteins due to proper glycosylation • Good secretion capability	• Slow growth compared to prokaryotes • More complex culture conditions than prokaryotes • Expensive growth media
Plant / Algal 	• Fundamental plant biology studies • Biofuel production	• Low growth cost • Can compartmentalize and accumulate protein in natural organs	• Technologies less developed compared to other systems
Cell-free 	• Toxic proteins • Fundamental studies including isotope labeling or unnatural amino acid incorporation	• Fast expression and purification process • Simple format	• Limited scalability • Expensive

1.5 Thesis outline

The overall aim of this thesis is to explore the genetic factors that contribute to gene expression and try to normalize or predict them. Eventually, these insights may result in genetic constructs that can be used in protein production platforms with reliable, efficient and cost-effective protein production. In the course of this research, attempts were made to gain fundamental insights through so-called randomization approaches in which suspected influential regions were partially or completely randomized. The effects of these regions were then studied by analysing and sequencing the randomized genetic libraries. Each data point and associated sequence can then be used to discover correlations to pinpoint the important sub-region and mode of action. In all experiments, fluorescent proteins were used as a case study as they are easy to quantify by measuring their fluorescence, which correlates well with functional protein production. Expression studies were done in *Escherichia coli* as this is an easy genetically accessible bacterium, allowing for the transformation and screening of a large number of genetic variants in a single experiment. With this overall approach, the 3 main contributing regions of the mRNA have been studied: the 5′ untranslated region (5′UTR), the coding sequence (CDS) and the 3′ untranslated region (3′UTR).

Firstly, in **Chapter** 2, a detailed overview of the then understanding of genetic design principles is provided. The role of both protein-coding and non-coding sequences are discussed for both prokaryotes and eukaryotes. This overview highlights many interconnected factors already known that contribute to overall protein production. Some of these genetic factors can already reliably be designed *a priori*, but the fundamentals of many other factors are still unknown, limiting synthetic designs. Finally, studies were highlighted that try to tackle the fundamental principles behind these interconnected genetic factors by generating large numbers of sequence variants and analyzing their effect on protein production.

In **Chapter** 3, the potential of a bicistronic design as a standardized alternative for the 5′UTR was explored. Secondary structures involving the RBS in the 5′UTR of an mRNA can be detrimental for protein production as they can prevent translation initiation by blocking recruitment of the ribosome. This phenomenon can partly explain why (while using the same expression vector) some proteins express very well in combination with a certain 5′UTR whereas others result in undetectable levels of protein. A bicistronic design is a naturally occurring genetic system that couples the expression of a gene to the expression of another gene. This system results in more predictable expression levels for the gene-of-interest. This is particularly important when studying codon usage as different codon sequences can each form secondary structures with the RBS with different levels of strength. A bicistronic design can drastically change the relative performance of different codon optimization algorithms. In a particular case, the performance of a codon optimization algorithm seemed very bad as the expression was very low. However, upon applying a bicistronic design, the expression levels improved drastically, showing that it was not the codon usage that was limiting, but rather the secondary structure that formed between the RBS and the codons. This was proven by disrupting the inhibitory secondary structure via a synonymous point mutation. Manual removal of the secondary structure yielded expression levels similar to the levels when using a bicistronic design.

Chapter 4 describes the major part of this thesis where a novel way to generate codon random genes is described which were used to develop a predictive algorithm. A large gene library with synonymous codon sequences was generated, which resulted in the exact same red fluorescent protein (RFP) but at different expression levels. Of this library, 1459 clones were characterized in detail: the DNA sequence of the *rfp* gene, and the fluorescence level of corresponding cells as a measure of protein production. We compared some of the better performing variants against modern and often used codon optimization algorithms and, interestingly, observed increased expression. This highlights that there is room for improving current optimization algorithms. This is probably due to incomplete knowledge of fundamental gene expression principles, as mentioned above. The expression levels and sequence data of the selected codon sequences were used to develop a machine learning algorithm that could predict the expression very well with a Pearson correlation of 0.803. We further used our algorithm to screen for hotspots in the gene by using a sliding window approach. This showed that most of the expression variation can be explained by the codons 2-9 (bases 5-25) of the coding sequence. This work provides strong evidence that a key factor for tuning protein production is the mRNA secondary structure between the RBS and the 5′ end of the CDS. This implies that codon usage is very important in that same region of the CDS (codons 2-9), but much less in the rest of the CDS.

In **Chapter** 5, the effects of gene placement in operons was studied. Next to earlier reported forward translational coupling, we provide evidence for a different, reverse coupling in which the rate of translation of a downstream gene influences that of the upstream gene. Additionally, the extreme effect of a Rho-independent transcriptional terminator on expression is highlighted. Finally, the untranslated region between the stop codon and the terminator was studied. A library of 30 nucleotide randomized sequences was generated, and a major effect on protein production was observed. Surprisingly, this untranslated region seems to act on translation in a CDS-independent way. By analyzing three different coding sequences (GFP, RFP, LacZ), it was observed that good performing sequences lead to high expression regardless of the CDS they are placed behind, and vice versa. This sequence region can act as an additional expression tuning device in *E. coli*, and most likely in other prokaryotes.

In **Chapter** 6, a summary of this thesis is provided. Finally, the overall results of this thesis are discussed and the results of pilot experiments for potential future research lines are presented. An outlook is presented where sequence randomization and selection for the best performing variants is presented as an alternative to *a priori* design.

Chapter 2

The ongoing guest to crack the genetic code for protein production

Adapted from:

Nieuwkoop, T., Finger-Bou, M., van der Oost, J., & Claassens, N. J. (2020). The Ongoing Quest to Crack the Genetic Code for Protein Production. *Molecular Cell*.

Abstract

Understanding the genetic design principles that determine protein production remains a major challenge. Although the key principles of gene expression were discovered 50 years ago, additional factors are still being uncovered. Both protein-coding and non-coding sequences harbour elements that collectively influence the efficiency of protein production by modulating transcription, mRNA decay, and translation. The influences of many contributing elements are intertwined, which complicates a full understanding of the individual factors. In natural genes, a functional balance between these factors has been obtained in the course of evolution, whereas for genetic-engineering projects, our incomplete understanding still limits the optimal design of synthetic genes. However, notable advances have recently been made, supported by high-throughput analysis of synthetic gene libraries as well as by state-of-the-art biomolecular techniques. We discuss here how these advances further strengthen understanding of the gene expression process and how they can be harnessed to optimize protein production.

2.1 Introduction

The biosynthesis of proteins is one of the core processes in living cells, as well as in many biotechnological applications. It has already been 50 years since Francis Crick proposed the central dogma of molecular biology (Crick, 1970), explaining how DNA is transcribed to mRNA, which is then translated to protein. A characteristic feature of the conversion of the information stored in the nucleotide building blocks of DNA and mRNA into the amino acid building blocks of proteins is the redundancy in the number of codons on the nucleotide level. Although there are 64 unique codons (nucleotide triplets), only 20 different amino acids make up proteins in most organisms. This redundancy gives astronomical numbers of codon combinations to encode the same amino acid sequence, e.g., the medium-size green fluorescent protein (GFP, 238 amino acids) can be encoded by 3 x 10^{110} different open reading frames (ORFs).

However, different sequences encoding an identical protein sequence can lead to dramatic variations in protein production levels, and sometimes even lead to differences in protein folding and functionality(Buhr et al., 2016; Kim et al., 2015; Zhou et al., 2013) (Figure 2.1). Apart from ORFs, non-coding regions with potential regulatory functions, such as promoters and untranslated regions (UTRs; Figure 2.1), add a vast sequence space. As the design principles of both the coding and non-coding sequences are only partly known, the design of synthetic genes for expression is still a major challenge.

Already, since the early days of gene sequencing in the 1980s, a bias has been recognized in the codon usage of highly expressed native genes; particular synonymous codons (i.e., different codons encoding the same amino acid) were observed to be used more frequently than others. This notion led to the formulation of the Codon Adaption Index (CAI) (Sharp and Li, 1987), and it was postulated that the codon bias within highly expressed genes allowed for more-efficient translation. An underlying hypothesis to this observation is that the (amino-acid-charged) cognate tRNAs for these frequent codons are more abundant and that they are more-efficient decoders during ribosomal protein biosynthesis (Ikemura, 1985; Reis, Wernisch, and Savva, 2003). In recent decades, the advent of high-throughput sequencing technologies has revealed more codon usage signatures varying across organisms, tissue types, and genes (Hanson and Coller, 2018)

Following these observations, several types of codon bias and mechanistic explanations were introduced (Quax et al., 2015). The current view on codon usage

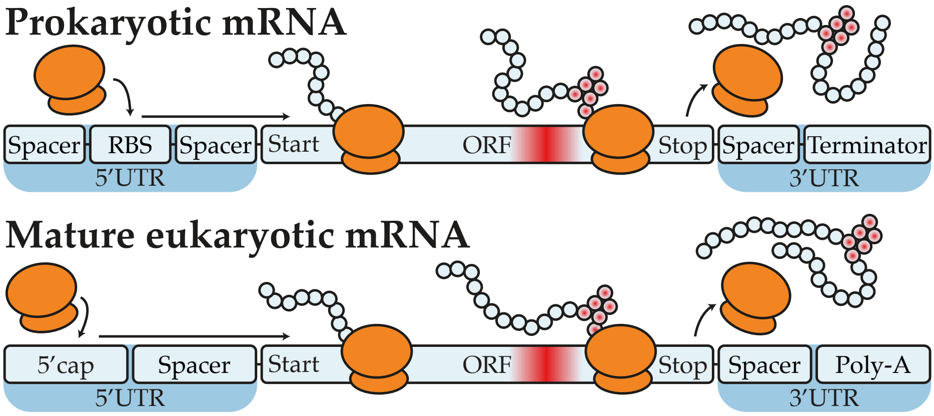


FIGURE 2.1: **Schematic Overview of a Prokaryotic and Mature Eukaryotic mRNA Being Translated by Ribosomes (Orange).** RBS, ribosome binding site; ORF, open reading frame; 5′/3′UTR, 5′/3′ untranslated region. The co-translational folding phenomenon is indicated with a red gradient in the mRNA and the associated amino acids.

is that it is related to a complexity of factors. The weight of those factors varies depending on the context, which includes the type of organism, tissue, or compartment; physiological control (e.g., pathway or growth phase); or even the position within an ORF (Hanson and Coller, 2018; Quax et al., 2015). It became clear that the notion of frequent versus rare codons, similar to good versus bad codons for protein production, is an oversimplification of biological reality. Consequently, codon optimization algorithms, which are all based on simplified assumptions and codon indices (Bourret, Alizon, and Bravo, 2019), cannot warrant successful heterologous protein production (Parret, Besir, and Meijers, 2016). Because codon choices are related to diverse mechanisms and regulatory processes, we prefer to use the term "codon optimality" only when a range of factors acting at different levels of the expression process have been taken into account.

A couple of years ago, we reviewed the effect of codon usage within the ORF on expression (Quax et al., 2015). Impressive advances have been made in the field since then, because, on one hand, of the technical advances, including high-throughput analyses of large synthetic gene libraries (Cambray, Guimaraes, and Arkin, 2018) and, on the other hand, because of innovative molecular biology approaches that unravelled additional details of transcription, translation, and protein folding (Buhr et al., 2016; Buschauer et al., 2020; Kim et al., 2015). These studies contributed to a further understanding of some of the factors involved and have also revealed relevant interactions among them.

Here, we provide a timely overview of the field of gene expression, discussing relevant features both in the regulation of non-coding regions and in ORFs. As transcription, mRNA decay, and translation (initiation and elongation) all have important roles in controlling protein production, we discuss all these stages (Figure 2.2). Furthermore, we highlight key controversies and knowledge gaps in the field and propose potential avenues to resolve these. Lastly, we discuss how our relatively poor understanding of optimal gene designs is a major limitation for biotechnology and synthetic biology. We examine how emerging tools and approaches can aid in overcoming challenges for engineering protein production.

2.2 Transcription and mRNA Decay

2.2.1 Transcription Initiation

The first step in protein production is the transcription of DNA to mRNA by RNA polymerase (RNAP). Synthesis rates of mRNA are mediated by the binding affinity of RNAPs and related transcriptional factors with the promoter sequences; other factors, such as chromatin structures in eukaryotes, also have a role (Lenstra et al., 2016). In addition, the transition from transcription initiation to transcription elongation is important in determining mRNA synthesis rates. After the RNAP is bound, DNA is unwound, and an open complex is formed. During the open complex configuration, the first short RNA stretch is transcribed, and then, the RNAP either moves on to transcribe the full mRNA (promoter escape) or the initiation is aborted. Several promoter sequence features, for example, the length and nucleotides in the bacterial discriminator region (±4–7 bp upstream of the transcription start), determine the efficiency of the promoter escape (Henderson et al., 2017; Winkelman et al., 2016). Promoter sequence regions, as well as transcription initiation and elongation factors involved in promoter escape, are reviewed in more detail elsewhere (Lee and Borukhov, 2016; Wade and Struhl, 2008). Although most of the key principles of transcription initiation and promoter escape are known, models to predict promoter strengths from sequences are still under development.

Recently, several groups investigated promoter properties and design constraints by expressing some reporter genes from libraries with randomized promoter sequences. Some studies in *Escherichia coli* reported that, of all fully randomized promoter sequences, 7%-10% resulted in detectable expression (Urtecho

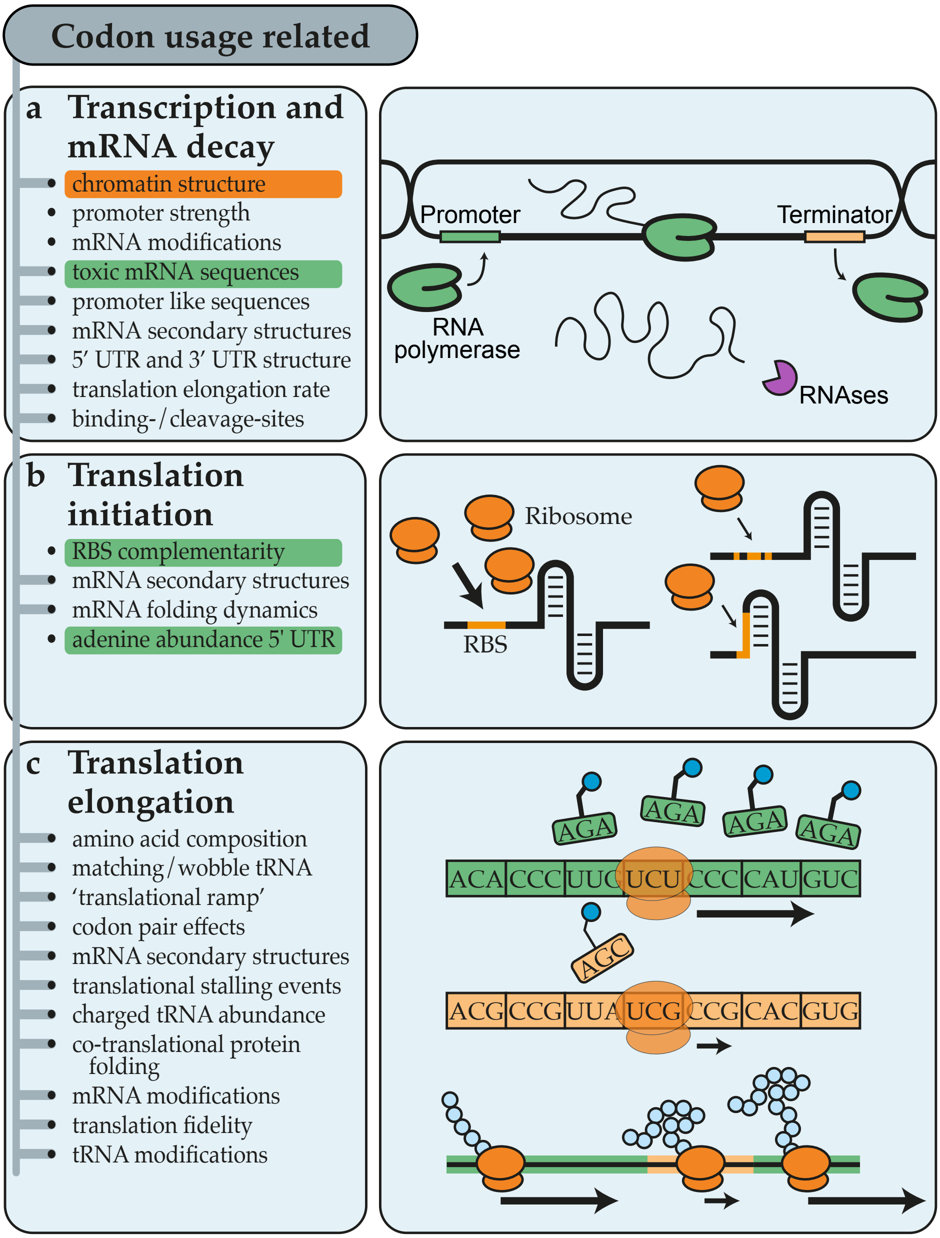

Codon usage related
a Transcription and mRNA decay
chromatin structure
promoter strength
mRNA modifications
toxic mRNA sequences
promoter like sequences
mRNA secondary structures
5′ UTR and 3′ UTR structure
translation elongation rate
binding-/cleavage-sites
Promoter
Terminator
RNA polymerase
RNAses
b Translation initiation
RBS complementarity
mRNA secondary structures
mRNA folding dynamics
adenine abundance 5′ UTR
Ribosome
RBS
c Translation elongation
amino acid composition
matching/wobble tRNA
'translational ramp'
codon pair effects
mRNA secondary structures
translational stalling events
charged tRNA abundance
co-translational protein folding
mRNA modifications
translation fidelity
tRNA modifications
AGA
AGA
AGA
AGA
ACA CCC UUC UCU CCC CAU GUC
AGC
ACG CCG UUA UCG CCG CAC GUG

FIGURE 2.2: **An Overview of Reported Factors Involved in Protein Production. (a–c)** Factors at the level of **(a)** transcription and mRNA decay, **(b)** translation initiation, and **(c)** translation elongation. Factors that can be related to codon usage are connected by the gray bar. Factors that have only been well-described in eukaryotes (orange) or prokaryotes (green) are highlighted; other factors have been observed in both domains of life.

et al., 2020; Yona, Alm, and Gore, 2018). Furthermore, it was found during laboratory evolution of random sequences in *E. coli* that 60% of those sequences became functional promoters with only one mutation (Yona, Alm, and Gore, 2018). Functional promoters in *E. coli* were generally observed to have at least a canonical -10 or -35 motif for binding the RNAP-sigma subunit, which occurs relatively frequently in DNA sequences by chance. Another study randomized the yeast -90 to -170 promoter region, whereas the consensus TATA region was kept constant, which resulted in detectable expression for 83% of the sequences (Boer et al., 2020).

The increasing data on characterized (random) promoters has also been used to create predictive models. Such *in silico* predictions have been successful for predicting promoter strengths of yeast, by modelling the transcription factor binding sites and their accessibility (Boer et al., 2020; Levo et al., 2017). However, the generation of predictive models for *E. coli* based on a set of fully randomized and native promoters by machine learning was still unsuccessful (Urtecho et al., 2020). This may be explained by the diverse sigma-factor-type promoters that are included in the training set. A previous study that performed machine learning and regression only on sigma-70 "household" promoters in *E. coli* did result in good predictive models (Urtecho et al., 2019).

The relatively high chance for random sequences to act as a promoter may also create "accidental promoters" in natural or synthetic sequences, which can cause transcriptional burdens and other distortions when they occur in undesired loci. Relatedly, an evolutionary selection against promoter-like sequences was observed within ORFs in *E. coli* (Yona, Alm, and Gore, 2018). Promoters within ORFs may, however, also serve functional roles occasionally; it has been proposed that promoters in the reverse sequence of ORFs can produce antisense RNAs to downregulate protein production (Brophy and Voigt, 2016; Urtecho et al., 2020).

Apart from the influence of promoter regions on transcription, it was observed in some eukaryotes that the codon or nucleotide usage within an ORF might also affect transcription rates (Fu et al., 2018; Newman et al., 2016; Zhou et al., 2016).

Proposed mechanisms through which nucleotide composition or codons could modulate transcriptional activity are related to histone modifications or the influence of GC-content on transcription elongation rates.

2.2.2 mRNA Decay

All cells harbour several endo- and exo-ribonucleases that are involved in degrading mRNA, providing additional control over mRNA levels and protein production (Schmid and Jensen, 2018). Furthermore, ribonucleases can clean up non-functional RNAs, e.g., from accidental transcription. The dynamics between mRNA transcription and mRNA decay result in a wide range of mRNA half-lives, serving as one of the key factors for protein production (Boël et al., 2016; Lahtvee et al., 2017; Presnyak et al., 2015).

One of the factors modulating mRNA stability is the presence of structural elements in their untranslated regions. Secondary structures and sequences of UTRs can influence mRNA decay rates, especially in bacteria (Mohanty and Kushner, 2016). Recently an increasing number of studies demonstrated the important role of the 3′UTR region in controlling mRNA decay (Menendez-Gil et al., 2020; Zhao et al., 2018). For the , it is harder to determine the effect of the sequence itself on mRNA stability because that region also has a key effect on translation initiation. In eukaryotes, 5′ caps and 3′ poly-A tails (Figure 2.1) are the primary features of the UTR regions that protect mRNAs from degradation (Mugridge, Coller, and Gross, 2018).

Diverse, alternative polyadenylation mechanisms in eukaryotes are activated by different signals in 3′UTR sequences and lead to differing poly-A tails and 3′UTR lengths; this region is highly interactive with RNA binding proteins, microRNA and long noncoding RNAs. These interactions and the 3′UTR length influence mRNA stability and decay, but also influence mRNA translation, as extensively reviewed elsewhere (Tian and Manley, 2017).

In the past decades, it has been suggested that the translation process may influence mRNA stability in yeast, as reviewed previously (Hanson and Coller, 2018). More recently, this connection gained additional attention in extensive studies in a range of eukaryotes, which all clearly demonstrated a positive correlation between the presence of certain codons in ORFs and the stability of the corresponding mRNAs (Bazzini et al., 2016; Burow et al., 2018; Forrest et al., 2020; Harigaya and Parker, 2016; Hia et al., 2019; Jeacock, Faria, and Horn, 2018; Mishima and Tomari, 2016; Narula et al., 2019; Freitas Nascimento et al., 2018;

Presnyak et al., 2015). In particular, specific codons are observed to be more abundant in mRNAs with a longer half-life. This observation was captured by a newly proposed codon index, the codon stability coefficient (CSC), which can be calculated for each codon as the correlation coefficient between the codon frequency in transcripts and their mRNA half-life (Presnyak et al., 2015) (Figure 2.3a). In several studies, it was found that this coefficient correlates moderately with the tRNA availability index (tAI). The latter index is based on the gene copy number of tRNAs available to decode a certain codon (Presnyak et al., 2015; Reis, Wernisch, and Savva, 2003). The observation that codons leading to high mRNA stability seem related to more-abundant tRNAs, remarkably suggests that the translational process may influence the stability of mRNAs. This was further supported by experiments that compared the mRNA stability with and without blocking the translation process (Bazzini et al., 2016; Wu et al., 2019). These experiments showed that when translation is inhibited, the mRNA half-life times are reduced, especially for transcripts with high "codon optimality."

On top of codon identity, a link is also suggested between amino acid identity and mRNA decay. A few amino acids are also specifically correlated to more or less stable mRNAs (Bazzini et al., 2016; Forrest et al., 2020; Narula et al., 2019; Wu et al., 2019). It is hypothesized that for these amino acids' higher or lower intracellular concentrations influence the amount of available tRNAs for translating those amino acids and hence influence translation elongation rates and consequently mRNA stability. In summary, several lines of evidence suggest that faster translation elongation leads to higher mRNA stability.

A potential molecular mechanism connecting translation elongation rates to mRNA decay has recently been unravelled (Figure 2.3B). Clear evidence was found in yeast that the de-adenylating Ccr4-Not complex directly interacts with ribosomes that are not loaded with a new tRNA in their A-site (Buschauer et al., 2020). Hence, this complex can sense slow-moving ribosomes and then triggers de-adenylating of the poly-A tail; after which, the RNA helicase Dhh1p activates de-capping, eventually resulting in mRNA decay (Mishima and Tomari, 2016; Radhakrishnan et al., 2016; Webster et al., 2018).

A link between codon usage and mRNA stability was also suggested for the bacterium *E. coli* to have a major role in protein production efficiency (Boël et al., 2016). This study focused on expression data from a large set of plasmid-encoded heterologous genes transcribed by T7 RNAP. So far, no genome-wide analyses are available on such correlations in bacteria for native gene expression.

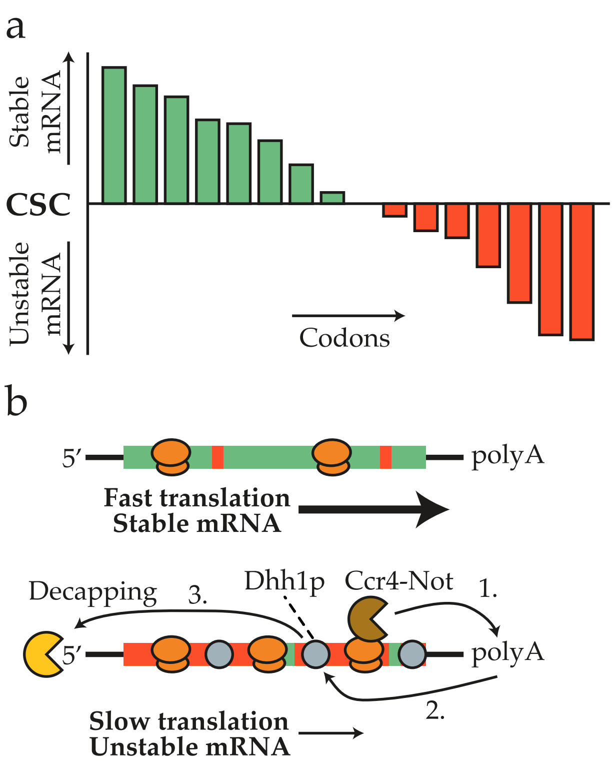


FIGURE 2.3: **Codon Usage and Translation Elongation Are Related to mRNA Stability in Several Eukaryotes. (a)** A schematic representation of a codon-stabilization coefficient (CSC) plot, based on recent studies in several eukaryotes, e.g., Presnyak et al., 2015. Bars for each codon represent the correlation between the codon frequency in the transcripts and the half-life of the transcripts. Positive correlations (green) indicate codons that are more abundant in mRNAs with a longer half-life time, whereas negatively correlated codons (red) are overrepresented in less-stable mRNAs. For illustrative purposes, only a few codons are depicted; in a real plot, the CSC value for all 61 amino-acid-encoding codons would be shown. **(b)** mRNAs with more codons with a high, positive CSC value (green) are observed to be translated faster by the ribosomes because, for example, those codons have more abundant cognate tRNAs. In the eukaryotic model organism yeast, a molecular mechanism has been elucidated that can explain the connection between slowly translated mRNAs and mRNA decay rates. The de-adenylating Ccr4-Not complex can directly interact with ribosomes that are not loaded with a new tRNA in their A-site (Buschauer et al., 2020). Likely, this complex senses slow-moving ribosomes and then triggers de-adenylating of the poly-A tail, and next the RNA helicase Dhh1p activates decapping and subsequent mRNA decay.

In relation to that, it is interesting to note here that recent structural studies in *E. coli* and *Mycoplasma pneumonia* clearly show that the RNAP complex can be linked to ribosomes in a so-called expressome, which leads to the coupling of transcription elongation to the translation process (O'Reilly et al., 2020). However, it was also recently reported that this coupling is not present in all bacteria because it was demonstrated in *Bacillus subtilis* that its RNAP moves faster than its ribosomes, in so-called runaway transcription (Johnson et al., 2020). The consequences of the presence and absence of this mechanism in different bacteria for the influence of codon usage and translation elongation on transcription deserve further analysis.

Lastly, another mRNA-mediated mechanism was discovered in *E. coli*, in which specific heterologous sequences of the mRNA appear to be toxic to the bacterial cells. It is not uncommon that the expression of heterologous proteins causes growth retardation in the expressing host, usually related to a protein production burden. However, a recent study surprisingly demonstrates that the growth retardation for specific heterologous mRNAs still happens when translation is blocked (Mittal et al., 2018). It is hypothesized that specific mRNA secondary structures cause toxic effects in the cell via a yet unknown mechanism.

Overall, our understanding of control mechanisms that determine mRNA concentrations is increasing. It is clear that mRNA abundance is affecting the downstream translational process and, remarkably, also vice versa translational processes seem to exert control on mRNA levels.

2.3 Translation Initiation

For transcripts to be translated into protein, ribosomes need to associate with the 5'UTR of the mRNA and start translating the ORF from the start codon. The translation initiation process is considered one of the most influential steps in translation efficiency.

In prokaryotes, it is generally assumed that translation initiation begins when the 30S ribosomal subunit recognizes a ribosome binding site (RBS) in the 5'UTR. The RBS usually contains a Shine-Dalgarno (SD) sequence, which has high complementarity to the 3' end of the 16S rRNA of the 30S ribosomal subunit, the so-called anti-Shine-Dalgarno sequence (aSD) (Shine and Dalgarno, 1974). In eukaryotes, the ribosome binds the 5' cap or an internal ribosome entry site (IRES) and usually translation initiation is further controlled by a Kozak sequence

(Kozak, 1981), a motif surrounding the start codon with a relatively high abundance of adenines (Leppek, Das, and Barna, 2018). However, because most recent studies on translation initiation used *E. coli* as a model, we mostly discuss prokaryotic translation initiation. For detailed insights on translation initiation and the 5'UTR in eukaryotes, we refer to other recent reviews (Leppek, Das, and Barna, 2018; De Nijs, De Maeseneire, and Soetaert, 2020).

Numerous studies, mostly investigating heterologous protein production in *E. coli*, have found that strong mRNA secondary structures around the RBS/SD region severely hamper translation initiation (Boël et al., 2016; Cambray, Guimaraes, and Arkin, 2018; Goodman, Church, and Kosuri, 2013; Kudla et al., 2009). The mRNA folding in this region is also regularly observed to be influenced by the codon usage at the start of the ORF. A recent study aimed to quantify the influence of mRNA secondary structures more accurately by designing strong RNA hairpins in the 5'UTR region of a reporter protein. Although secondary structures located far from the SD only result in less than 2-fold repression of translation, secondary structures close to the SD were shown to repress translation more than 100-fold; the repression levels are proportional to the free energy needed to unfold the RNA hairpins (Espah Borujeni et al., 2017). Furthermore, a study that introduced synonymous codon mutations throughout ORFs of two native *E. coli* genes revealed that, especially mutations leading to relatively strong, predicted mRNA secondary structures that include the RBS, result in significantly decreased protein production levels (Bhattacharyya et al., 2018).

Although most studies base their mRNA structure predictions on *in silico* folding energy models, some recent studies have applied transcriptome-wide *in vivo* experiments to determine mRNA secondary structures. Experimental high-throughput measurements of mRNA secondary structures can be performed by cell-permeable chemicals that react selectively with non-paired RNA bases, e.g., SHAPE probes that acylate 2' hydroxyl groups of unpaired nucleotides (SHAPE-MaP) (Siegfried et al., 2014) or dimethyl sulfate that modifies unpaired adenine and cytosine residues (DMS-seq) (Rouskin et al., 2014). As the next step, cDNA is generated from the chemically modified RNAs, and next-generation DNA sequencing allows for mapping of the modifications in non-structures regions and, hence, allows the elucidation of non-structured and structured mRNA regions. One of these studies, based on SHAPE-MaP in *E. coli*, demonstrated that the translation efficiency of native genes is, in large part (40%), determined by mRNA structures covering the RBS (Mustoe et al., 2018).

The improved resolution of mRNA structure measurements also allowed the study of two alternative models for translation initiation: the equilibrium model and the kinetic model. In the equilibrium model, the ribosome, once bound, remains and creates a new equilibrium mRNA secondary structure. In the kinetic model, however, there is a continuous competition between the unfolding and refolding of the mRNA and association and dissociation of the ribosome. Experimental data, as well as a theoretical biophysical approach, now suggest the kinetic model best explains translation initiation in *E. coli* (Espah Borujeni and Salis, 2016; Mustoe et al., 2018). This also allows for "ribosome drafting" in some highly translated mRNAs, a mechanism in which successive ribosomes bind an mRNA faster than the mRNA can refold.

In contrast with the ribosome drafting mechanism, in eukaryotes, it was observed that ribosome clearance around the translation initiation site is required for high-expressing genes. It is suggested that codons directly after the start codon need to mediate relatively fast translation elongation to free up space for the next ribosome to initiate translation (Chu et al., 2014).

Although it is generally accepted that SD-aSD interaction is the main player involved in prokaryotic ribosome loading, new findings hint at alternative mechanisms regulating ribosome recruitment and translation initiation. Several bacterial species, for example, *Flavobacterium johnsoniae*, naturally lack SD sequences. In this species, it was observed that at some key nucleotide positions upstream of the start codon (-3, -6, -13, and -23), the presence of adenine nucleotides is a positive determinant for translation initiation (Baez et al., 2019). The molecular basis for this observation is currently not known but, as the authors state, it seems reminiscent of the eukaryotic Kozak sequence, which also shows a preference for adenine at position -3. Furthermore, some recent *E. coli* studies on native and reporter gene expression report an enrichment in adenines at sites mostly upstream, or shortly downstream of the start codon for well-expressed genes (Komarova et al., 2020; Saito, Green, and Buskirk, 2020). It was demonstrated experimentally that these A-rich sequences contribute to the identification of translational start sites, suggesting that these adenines could be highly conserved as an alternative mechanism for start site selection in bacteria (Saito, Green, and Buskirk, 2020).

2.4 Translation Elongation

2.4.1 Codon Usage and Translation Rates

After successful initiation, ribosomes continue with translation elongation, i.e., the sequential decoding of the codons of the mRNA to synthesize the corresponding amino acid sequence. The effect of codon usage during translation elongation has been extensively studied by multiple methods, however, often leading to contrasting conclusions. A popular hypothesis is that codon usage controls the speed of ribosomal translation elongation. The underlying assumption is that translating ribosomes slow down when they encounter "sub-optimal" codons, e.g., codons that are decoded by less-abundant (amino-acid-loaded) cognate tRNAs or by lower-affinity-matching tRNAs through wobble base-pairing.

A decade ago, the ribosome profiling technique was developed to monitor translation elongation rates in a high-throughput manner (Ingolia et al., 2009). This approach is based on the high-throughput sequencing of ribosome-protected mRNA fragments, providing a snapshot of ribosome density throughout the transcriptome. Initially, differences in experimental ribosome profiling protocols and subsequent data analysis led to conflicting conclusions on whether translation elongation speeds are influenced by codon usage or not (Charneski and Hurst, 2013; Gardin et al., 2014; Li et al., 2014; Quax et al., 2013). However, in recent years ribosome profiling protocols and data analysis were refined, e.g., by the use of flash freezing to stall translation, instead of the use of cycloheximide (Weinberg et al., 2016). Improved protocols led to a better consensus that codon usage may influence the translation elongation speed, but that this effect is rather weak and that a multitude of other factors are also involved (Hanson and Coller, 2018).

Recently, more-sensitive approaches using cell-free translation systems (Buhr et al., 2016; Yu et al., 2015) and *in vivo* imaging of nascent polypeptide synthesis (Chekulaeva and Landthaler, 2016; Yan et al., 2016) have been established. These methods all confirmed that heterologous mRNAs with "optimal" codon usage are translated faster. However, these studies monitored the strong contrast between synthetic genes that were designed to have almost only optimal codons with non-optimized genes. Within natural genes, which often have fluctuating use of optimal codons along the ORF, translational speed differences are generally more subtle.

It was also demonstrated for eukaryotic translation, both *in vivo* and *in vitro*, that rare codons sometimes not only slow down translation, but they can even stall part of the elongating ribosomes, leading to premature translation termination (Yang et al., 2019; Yu et al., 2015; Zhao, Yu, and Liu, 2017).

2.4.2 Does an mRNA Secondary Structure Influence Translation Elongation?

Besides the influence of codon usage on translation speed, the mRNA secondary structure within an ORF was also suggested as influencing translation elongation. However, until recently, it was hard to verify that hypothesis because only rough *in silico* predictions of mRNA folding energy were available to estimate mRNA structures. However, the aforementioned development of several experimental protocols allows for probing RNA structure *in vivo* at a transcriptome-wide scale. Two studies in this field used different methods to both reach the conclusion that translating ribosomes in *E. coli* dissolve RNA secondary structures (Burkhardt et al., 2017; Mustoe et al., 2018), which is in line with the demonstration that the *E. coli* ribosome exhibits helicase activity (Takyar, Hickerson, and Noller, 2005).

Apart from that finding, the DMS-seq analysis by Burkhardt et al., 2017 reported a strong correlation between mRNA secondary structures in an ORF and its translation elongation efficiency, suggesting that at least some of those structures can still be an obstacle for translating ribosomes. In contrast, the SHAPE-MaP analysis by Mustoe et al., 2018 could not confirm that correlation. Hence, despite advances in *in vivo* RNA structure mapping, it remains unclear to what extent mRNA structures influence translation-elongation rates. Refinement and application of these methods throughout multiple organisms are required to clarify this matter.

2.4.3 Co-translation Folding Mediated by the ORF Sequence

For a few specific proteins, single-molecule approaches have been used to accurately monitor translation elongation rates and related co-translational protein-folding processes. In some cases, it was clearly shown that the slow-down of translation elongation is crucial to facilitate proper co-translational folding of the nascent protein (Buhr et al., 2016; Kim et al., 2015).

Similarly, it has been demonstrated *in vivo* for some eukaryotes that codon usage is crucial for the folding and functionality of some circadian clock proteins, especially for the unstructured domains of these proteins. When the sub-optimal codon usage in unstructured regions of these circadian clock genes, as well as in a luciferase reporter gene, was changed to a more-optimal codon usage, the *in vivo* functionality of these proteins was compromised (Fu et al., 2016; Yu et al., 2015; Zhou et al., 2013; Zhou et al., 2015). This folding hypothesis is further supported by broad bioinformatic analyses of genes from several organisms, based on which correlations are reported between less-optimal codons in unstructured regions in between more-structured protein domains (Pechmann and Frydman, 2013; Zhou et al., 2015). Despite the fact that these unstructured domains do not form defined structures (α helices or β sheets), they seem to have certain folds (e.g., coils) that can be essential for their functionality. These studies suggest that translation slows down to facilitate folding either of these unstructured domains themselves or at structural junctions between structured and unstructured domains.

However, a broader analysis of clusters of rare codons throughout many genomes in all domains of life challenges this observation of rare codons within unstructured domains (Chaney et al., 2017). That study, in fact, reports an enrichment of rare codons within structural domains, suggesting that translational slow-downs may be specifically relevant for the folding of smaller structural subelements. As an example, they show conservation of rare codon clusters for two proteins at the same "structural" positions throughout different organisms. Providing such comparative analyses for more proteins, as well as performing functional experiments on these, could strengthen the proof that sub-optimal codons are also relevant within structural protein domains.

Overall, there is clear case-based evidence on the effects of codon bias and translational speed on co-translational folding for some specific proteins. However, interpretation of these effects on a genome-wide scale is complicated, given the limited understanding of the genetic features determining the translational speed and the subjective definitions of optimal and non-optimal codons. Furthermore, determining the relevance of the coding sequence on protein folding is challenging, as it is currently not possible to experimentally determine protein structures or folding processes in a high-throughput manner.

2.4.4 Translation Effects at the Start of the ORF

Another frequently reported and heavily debated observation is the slower translation at the 5′ end of an ORF. Some evidence for this has been based on ribosome profiling data and the higher frequency of rare codons in the first part of the ORF (Tuller and Zur, 2015; Tuller et al., 2010). A main hypothetical explanation for the presence of a so-called translational ramp at that location is the distancing between ribosomes to prevent detrimental ribosomal collisions. Still, there are alternative explanations for the observed codon bias at the 5′ of ORFs. A key alternative hypothesis is that a strong selection against mRNA secondary structures at the 5′ end to facilitate translation initiation of highly expressed genes is more important than the selection pressure for well-translated codons in that region of the ORF.

Interestingly, several studies that randomized synonymous codons in *E. coli*, usually for GFP as a reporter protein, found strong correlations between protein production and reduced mRNA secondary structures around the 5′ end of the ORF (Goodman, Church, and Kosuri, 2013; Kelsic et al., 2016; Kudla et al., 2009).

A recent study tried to resolve the factors in the 5′ end of the ORF in a more systematic way by designing >200,000 different N-terminal tags for 32 codons, followed by a GFP reporter gene (Cambray, Guimaraes, and Arkin, 2018). Several factors were varied in the N-terminal library design, including the presence of different-strength translational ramps, as well as the presence of mRNA secondary structures at different positions. Although no correlation was detected between translational ramps and expression, that study did demonstrate a major role of mRNA structural elements in RBS availability and, consequently, in overall protein production. However, as the authors admit, the conclusion that the presence of a translational ramp could not be detected in that study might have been the result of non-optimal design. Although it remains unclear to what extent translations ramps influence expression levels, it was demonstrated recently that a ramp can decrease the resource costs of expression (Frumkin et al., 2017), likely by preventing ribosome jamming and translational abortion events (Tuller et al., 2010).

2.4.5 Other Factors Observed at the Translational Level

Apart from the effect of single codons on translational dynamics, it was observed previously that specific codon pairs might also influence translational processes (Buchan, 2006; Gutman and Hatfield, 1989). In yeast, ribosomal stalling has been reported for a small subset of codon pairs, mostly when they occur in a specific order (Gamble et al., 2016). Recently, a mechanistic explanation for that observation was found. It was determined that interactions of specific codons pairs with their tRNAs, mostly involving wobble-base pairing, induce certain conformational changes in the ribosomes that lead to stalling (Tesina et al., 2020).

The use of sub-optimal pairs of codons has also been proposed as a strategy to create live-attenuated viruses for vaccine development. However, there has been a lively debate about whether the decreased expression of those viruses in eukaryotic host cells should be attributed to suboptimal codon pairs or, alternatively, to sub-optimal dinucleotide pairs (Kunec and Osterrieder, 2016). A recent study that aimed to disentangle the effects of dinucleotide bias and codon-pair bias in virus attenuation concluded that sub-optimal codon pairs primarily caused the decreased translational efficiency (Groenke et al., 2020). That study shows that the influence of sub-optimal codon pairs can, at least partly, be related to decreased mRNA stability, in line with the previously discussed correlation between codon usage, translation efficiency, and mRNA stability in eukaryotes.

In bacteria, the presence of SD-like sequences within ORFs was previously suggested to result in a slow down of the translation-elongation process (Li, Oh, and Weissman, 2012). However, that observation was later toned down in a reevaluation of ribosome-profiling data, which concluded that SD-like sequences have little or no effect on translational pausing (Mohammad et al., 2016). Recently, a bioinformatical analysis studying the evolutionary conservation of those SD-like sequences in ORFs of several bacterial species, concluded that they are less conserved than would be expected by random chance (Hockenberry et al., 2018). This suggests a negative evolutionary selection against SD-like sequences, hinting at a potential decrease in fitness caused by the presence of those sequences within ORFs, possibly because they could induce mistranslation or erroneous frameshifting. In conclusion, it seems that SD-like sequences are not frequently used in nature because of detrimental by-effects on translation and that they do not have a major role in controlling translation elongation rates.

Another recent study has revealed an interesting effect of certain short amino acid motifs on translation elongation. That study focused on mutating codons at positions 3, 4, and 5 of a GFP reporter in *E. coli* and allowed non-synonymous mutations (Verma et al., 2019). They identified specific amino acid motifs at the start of the ORF that lead to high translation efficiency, independent of specific codons or mRNA structures. At the same time, they identified detrimental amino acid motifs in the 5′ region of the ORF, which can cause pausing of the translation and lead to increased translational abortion. This observation was explained by specific interactions of the nascent peptide motif with the ribosome exit tunnel that could lead to ribosomal stalling and drop-off.

There are more reports of specific peptide motifs that cause stalling or translational slowdown, likely via interactions in the ribosome exit tunnel. Motifs such as poly-proline sequences can slow down or stall translation in organisms throughout all domains of life (Huter et al., 2017; Wilson, Arenz, and Beckmann, 2016). In addition, it was observed in *E. coli* that four specific amino acid triplets completely stalled translation and were avoided within its proteome (Navon et al., 2016). It is good to realize that both in evolution and in synthetic biology approaches, the flexibility to evolve or design acceptable changes in amino acid sequences, without altering residues that are critical for protein functionality, may sometimes result in improved translation efficiency.

Furthermore, translational speed can be influenced by the modifications of mRNA and tRNAs. It is well established that the great diversity of tRNA modifications, especially modifications of ribonucleotides in anticodon regions, can have a major effect on translation rates and fidelity (Chou et al., 2017; Kimura, Srisuknimit, and Waldor, 2020; Nedialkova and Leidel, 2015). Recently, it was also observed that modifications of mRNA, e.g., N6-methyl-adenosine and N4-acetylcytidine, influence translation elongation and mRNA decay in both eukaryotes and bacteria (Arango et al., 2018; Choi et al., 2016; Zhao, Roundtree, and He, 2017).

2.4.6 Translational Fidelity versus Translation Rate and Translation Termination

Apart from governing translational speed, ORF sequence features such as codon usage have been postulated to govern translational fidelity. Even though support for this theory has been provided by bioinformatic analyses (Drummond and Wilke, 2008), only very recently has experimental evidence for this hypothesis been obtained. Using a "deep proteomics" approach, translational errors have been identified in the proteomes of *E. coli* and *Saccharomyces cerevisiae* (Mordret et al., 2019). That study revealed that translation errors are relatively abundant, occurring on average once every 1,000 amino acids. Transcriptional error rates occur much less frequently, at about 1 in 25,000 nucleotides (Traverse and Ochman, 2016).

Both the misloaded tRNAs and tRNA-codon mispairing can cause translation errors, but the latter error is more abundant. In that case, wrong amino acids are delivered by near-cognate tRNAs, which have only one mismatch between codon and anti-codon (Mordret et al., 2019). Interestingly, the effect of mistranslation events is probably reduced because the genetic code has evolved such that these near-cognate tRNAs often deliver amino acids with similar chemical properties. Some codons are more sensitive to mistranslation than others, and that pattern was relatively similar both in yeast and in *E. coli*, suggesting that evolutionarily conserved mechanisms or universal chemical interactions lead to occasional mistranslation.

The same study also demonstrated a negative correlation between translation speed and translation fidelity, suggesting a trade-off between optimizing coding sequences for translational speed and fidelity. This fidelity theory (slowdowns to reduce translational errors) is an interesting alternative explanation for the aforementioned occurrence of "slow" codons in structurally important regions, which, in many reports, is explained by the co-translational folding theory (Buhr et al., 2016; Kim et al., 2015).

Frameshifting during translation has an even bigger effect on protein function than amino acid misincorporation because the downstream sequence is completely mistranslated. However, the operation of ribosomes and their translation elongation factors seems to limit frameshifting. Recently, another mechanism for frameshift fidelity was observed in human cells (Wan et al., 2018). It was suggested that periodic pairing of certain "sticky codons" on the mRNA with complementary triplets in the rRNA, near the exit of the ribosomal mRNA channel, helps to prevent frameshifting. That conclusion was supported by the substitution of sticky codons by synonymous counterparts, which led to a 4-fold increase in frameshifting, as well as mutating the complementary triplet at the exit of the ribosomal mRNA channel, which also influenced the frameshifting rate. Finally, it seems that these sticky codons are naturally underrepresented in a non-coding frame in eukaryotic genomes, which may be to prevent accidental frameshifting (Wan et al., 2018). This mechanism deserves further analysis throughout different types of organisms and may cause certain codon preferences to limit frameshifting.

At the end of the translation-elongation process, the ribosome encounters a stop codon, and upon binding of a release factor (a protein mimic of a tRNA), the translation is ended, and the ribosome is released from the mRNA. However, in rare cases, translation read through happens, generally leading to the synthesis of non-functional proteins. If such a read-through event takes place, the ribosomes either encounter an in-frame stop codon within the 3'UTR or they get stalled at the end of the mRNA (Wilson, Arenz, and Beckmann, 2016). These read-through proteins are generally degraded co- or post-translationally (Arribere et al., 2016). Some organisms may prevent translational read through by using tandem stop codons, which are, for example, observed more frequently in the 3'UTR of ciliates (Fleming and Cavalcanti, 2019).

2.5 The Interactions between Different Factors

2.5.1 Cooperative and Counteracting Features

As discussed, distinct factors are involved in different steps of the gene-expression process, and they interact with each other in multiple ways. Some factors in the protein-production process act in a cooperative fashion. As a remarkable example of that, the translation-elongation efficiency and mRNA stability in eukaryotes have been demonstrated to be mechanistically linked, leading to positive feedback between translation elongation and mRNA stability (Buschauer et al., 2020; Radhakrishnan et al., 2016). However, other sequence features may also influence each other negatively. For example, a high-affinity SD sequence and well-translated codons in the 5' region of the ORF could form a base pair and, consequently, form undesired mRNA secondary structures that hamper efficient translation initiation. These counteracting and cooperative features complicate the evaluation of individual factors.

Several studies have attempted to reveal new factors and to disentangle their connections in recent years. Many of those studies applied randomized or systematically designed reporter gene-variant libraries of GFP in *E. coli*. (Cambray, Guimaraes, and Arkin, 2018; Frumkin et al., 2017; Goodman, Church, and Kosuri, 2013; Kudla et al., 2009). The consensus of those studies is that gene expression is significantly affected by strong (predicted) mRNA secondary structures in the 5′UTR and the 5′ region of the ORF. However, a large part of the variation in expression levels in those studies is explained by a range of other factors, and a substantial part of the observed fluctuations cannot be explained at all. Furthermore, it is not certain that those studies properly reflect features that are relevant to native genes. Nevertheless, a number of recent studies on native gene expression in *E. coli* also suggest that mRNA structures and associated RBS availability are key factors that determine the expression rate of natural genes (Kelsic et al., 2016; Mustoe et al., 2018).

A combination of different experimental approaches to study native gene expression was recently performed in yeast, integrating multiple omics data and measurements of mRNA and protein half-life times (Lahtvee et al., 2017). The latter is an often overlooked factor because proteins with shorter half-lives need to be translated at higher levels to sustain sufficient protein levels. That study found large differences in protein yield per mRNA, varying up to 400-fold among some proteins, suggesting an important role in the efficiency of the translation processes. However, when accounting for all proteins, translation-elongation efficiency only explained 15% of the protein abundance observed, whereas mRNA abundance was the most important explanatory factor for protein levels (explaining 61%). A large study on a diverse set of heterologous proteins in *E. coli* also reported mRNA abundance as the main predictor for protein abundance (Boël et al., 2016). However, it is important to realize that mRNA abundance can also be influenced by translation efficiency.

2.5.2 Influence of Gene Designs on Resource Consumption and Growth

An important, overarching aspect for protein production is the high metabolic costs associated with transcription and translation processes. Those additional costs include "materials", such as demands for ATP, nucleotides, and amino acids, but also the extra demand for the transcriptional and translation factors, such as RNAPs and ribosomes. There is an evolutionary pressure on the genome in general, and the architecture of genes and their regulation in particular, to reduce metabolic costs to optimize cellular fitness. Within synthetic-biology applications, the reduction of energy and resource requirements is of importance for gene design.

Hence, recent efforts studied growth parameters of microbial cells harbouring codon-variant libraries of reporter genes (e.g., GFP) or of a growth-essential gene. The relative fitness of different variants was recorded, either by measuring growth curves for individual strains or by performing competition experiments between them (Cambray, Guimaraes, and Arkin, 2018; Frumkin et al., 2017; Kelsic et al., 2016). One of the main conclusions is that, especially for highly expressed genes, a high level of protein produced per mRNA is a resource-efficient way for high expression. So, even though, in nature, high mRNA levels are typically correlated to high expression, boosting expression solely by high mRNA levels is not the best strategy. Extremely abundant mRNAs potentially imply excessively high transcription costs or may sequester excessive amounts of ribosomes from the limited pool. In contrast, we note that the strategy to keep mRNA levels low and, rather, to couple it to highly efficient translation can increase the cell-to-cell variability in mRNA and protein concentrations (Taniguchi et al., 2010). Thus, to achieve both high resource efficiency and low cell-to-cell expression variability, nature and synthetic biologists need to properly tune the translation efficiency per mRNA.

2.6 Biotechnological Challenges and Opportunities for Gene Design

Innovations in DNA synthesis and genetic engineering have tremendously accelerated the capacity to express synthetic genes. However, based on data from consortia aiming to resolve large numbers of protein structures, it is estimated that only about one-half of the attempts for heterologous protein production led to successful expression (Parret, Besir, and Meijers, 2016). In practice, in molecular biology and synthetic biology projects, the expression of synthetic genes regularly leads to sub-optimal production or problematic growth because of the excessive expression burdens.

2.6.1 Limitations of Codon Optimization Algorithms

Synthetic genes for heterologous protein production are typically designed with codon-optimization algorithms, which generally optimize a particular ORF, adapting it to a codon-usage index of the expression host (Parret, Besir, and Meijers, 2016). Those codon indices are frequently determined with either the codon usage within a set of highly expressed reference genes (e.g., CAI) or the tRNA copy numbers (e.g., tAI) in the host cell. Some academic and commercial algorithms also take alternative parameters into account, such as GC content and avoidance of certain regulatory motifs, such as SD sequences or repeats (Gould, Hendy, and Papamichail, 2014). Only a few algorithms additionally aim to minimize mRNA secondary structures (Gould, Hendy, and Papamichail, 2014), even though the folding in the translation-initiation region, certainly in prokaryotes, is a key determinant of expression. A promising exception is the novel 31C-FO algorithm, which aims to minimize mRNA folding of the 5'UTR and the first 48 bases of the ORF (Boël et al., 2016). At the same time, that algorithm optimizes codon usage by only including 31 codons that are correlated to high expression in *E. coli*. That algorithm was reported to lead to successful expression of several proteins by Boël et al., 2016 but has not been reported in other studies yet, and no easy tool for that algorithm is available so far.

Generally, the features involved in gene expression, individually or in concert with others, are still not understood in sufficient detail to compose robust optimization algorithms for relevant host organisms. Multi-parameter algorithms, such as EuGene or DNA-Tailor (D-Tailor) (Gould, Hendy, and Papamichail, 2014), typically leave the setting of specific objectives up to the users, which, in practice, is hard to decide upon, given the unknown weight of the different factors. Furthermore, it has been shown that a so-called design-of-experiments approach, which systematically varies multiple factors, is no guarantee for successful expression because not all relevant factors are known yet or are not known in sufficient detail (Cambray, Guimaraes, and Arkin, 2018).

In addition to expression levels, proper protein folding is important for the functional production of proteins. The accumulating evidence on the role of codon usage in protein folding led to several approaches that aimed to include the translation-speed landscape to accommodate the folding of structural elements. For example, codon-harmonization algorithms have been proposed to tackle this issue (Angov et al., 2008; Buhr et al., 2016). These algorithms have as their objective to copy the *native*-codon-usage landscape of a gene-of-interest (distribution of rare and frequent codons in the organism from which the gene originated natively) into a *heterologous*-codon-usage landscape (similar distribution of rare and frequent codons in the context of the expression host). However, codon harmonization does not always give the best expression levels in *E. coli* when comparing the production levels of codon-harmonized gene variants with native genes or CAI-codon-optimized genes for some membrane proteins (Claassens et al., 2017).

In some studies, sub-optimal codons or SD-like sequences have been included in ORFs to slow down translation in between structural domains, which was reported to improve protein solubility in a few cases (Hess et al., 2015; Vasquez et al., 2016). That, however, requires laborious, detailed studies to determine exactly the position and strength of the required translation pauses to optimize the folding of a specific protein. Furthermore, there is no full understanding yet on the role of the coding-sequence features for translational speed; this all restrains robust design approaches for proper folding of proteins.

In summary, improving heterologous protein production by codon-optimization algorithms often remains a trial-and-error approach. Success rates can be increased by testing multiple different codon-optimized variants, but that also increases experimental labour and costs.

2.6.2 UTR Optimization Strategies

In numerous studies, the 5′UTR has been identified as a critical region that determines translation-initiation efficiency in protein production. As discussed, few of the available codon-optimization algorithms take the 5′UTR into account and do not have integrated functionality to avoid detrimental mRNA structures in the translation-initiation region. Nonetheless, some specific tools have been developed to design optimized 5′UTR regions for bacterial protein production, which generally try to design 5′UTRs to have strong and accessible RBSs, taking into account the downstream ORF region. Hereto, these tools have used *in silico* mRNA folding energy calculations (Bonde et al., 2016; Jeschek, Gerngross, and Panke, 2016; Salis, Mirsky, and Voigt, 2009). Despite their wide use and relatively successful predictions, they still suffer from the limited reliability of *in silico* RNA structural predictions. Recently emerging experimental tools for measuring *in vivo* RNA folding may become helpful to assess the validity of computational predictions (Rouskin et al., 2014; Siegfried et al., 2014).

Alternatively, standardized 5′UTR modules have been employed for robust gene expression, for example, by using combinations of well-expressed 5′UTRs and N-terminal tags (Ki and Pack, 2020). In addition, bicistronic RBS modules have proven highly useful because these modules partly uncouple translation-initiation efficiencies from the ORF sequence (Cambray, Guimaraes, and Arkin, 2018; Mutalik et al., 2013a). These bicistronic design elements (BCDs) have been shown to allow for tuned and improved expression levels in *E. coli* and *Corynebacterium glutamicum* (Claassens et al., 2019; Nieuwkoop, Claassens, and Oost, 2019; Sun et al., 2020).

The initiation mechanisms in the 5′UTR in eukaryotes seem more diverse and complicated than do those for prokaryotes. However, recent studies have shown that the 5′UTR sequence has great potential for tuning the expression in eukaryotes, such as *S. cerevisiae* or Chinese hamster ovary-S (CHO-S) cells (Ding et al., 2018; Petersen et al., 2018; Weenink et al., 2018). These studies provided modular 5′UTRs designs that work relatively well with low-context dependence on the downstream ORF. One of the key factors that improve the performance of those 5′UTR is the reduction of mRNA secondary structures in that region.

The 3′UTR is less studied in relation to expression efficiency, but it has also been reported to influence mRNA stability and transcription termination efficiency, thereby modulating expression efficiency. Examples of 3′UTR engineering in bacteria are scarce, so far. For yeast and human cell lines, some short synthetic 3′UTR modules have been developed that relatively robustly increase expression for multiple genes throughout multiple species but also seem partly dependent on the upstream ORF sequence (Cheng et al., 2019; Curran et al., 2015).

An important part of the influence of 5′UTRs and 3′UTRs on protein production is explained by their roles in mRNA stability. An alternative, promising approach to improve mRNA stability for protein production, is through the circularization of mRNAs, which also occurs in nature. Synthetic circular mRNAs can, for example, be generated by harnessing the mechanism of self-splicing introns (Perriman and Ares, 1998; Wesselhoeft, Kowalski, and Anderson, 2018). A recent surge of research in this field showed promising applications for protein production driven by synthetic, circular mRNA transcripts in eukaryotes. Because canonical-eukaryotic translation initiation relies on the 5′ cap, alternative translation-initiation mechanisms, such as IRES or N^6-methyladenosine modifications, are required to ensure sufficient translation initiation in circular mRNAs. Furthermore, it has been proposed that the translation of circular mRNAs can be increased by creating an infinite ORF, by removing the stop codon of the ORF (Perriman and Ares, 1998); the same ribosomes will repeatedly translate the same sequence, leading to a multimeric protein, and individual functional proteins can be produced by introducing protease cleavage or self-cleavage sites in the polypeptide. A recent review elaborates in great detail on the developments of engineering of circular RNA (Costello et al., 2020).

2.6.3 Randomization, Smart Selection, and Machine Learning

The number of studies that randomly vary sequences in promoters, 5′UTRs, and the start of the ORF have steadily increased, mostly for GFP. This randomization approach may also be relevant for optimizing or fine-tuning the production of more biotechnologically relevant proteins (Figure 2.4). However, unlike expression levels of reporter proteins, levels of most proteins of interest are generally hard to screen with sufficient throughput from large randomized libraries. Still, some well-expressing modules identified in reporter-based screens, e.g., promoters or 5′UTR-N-terminal tag peptide combinations, have been used successfully for the optimized production of other proteins.

When randomly optimizing the coding sequence, or at the junction of the 5′UTR and coding sequence, novel approaches are required to screen for well-expressed gene variants in the case of non-reporter proteins. One simple approach is to fuse the protein of interest to a reporter protein, but such a fusion frequently distorts the function of the protein of interest. An alternative method, not based on a protein fusion, was recently established by translational coupling of the protein of interest to a selectable antibiotic-resistance reporter. This so-called TARSyn system was demonstrated for the high-throughput selection of optimized 5′UTR:ORF junctions for the expression of antibody proteins in *E. coli* (Rennig et al., 2018) (Figure 2.4). We consider the development of selection and screening systems of well-expressed "randomized" sequences to be a very promising avenue for further exploration.

Alternatively, data collected from large-scale randomization studies on reporter proteins or growth-selectable markers may help to generate better predictive algorithms (Figure 2.4). These large-scale data could serve as training sets for machine learning. Different types of machine learning can be employed to generate more reliable algorithms to improve the design of synthetic genes (Jongh et al., 2020).

A recent, innovative study that used machine learning focused on predicting the influence of different 5′UTR sequences in *E. coli* (Höllerer et al., 2020). The study developed an innovative reporter system, based on a recombinase protein, to quantify the expression from a large library of randomized 5′UTR sequences (Figure 2.4). At a certain expression level, that site-specific recombinase flips a DNA sequence, which is located directly next to the 5′UTR on the same plasmid. Subsequent, high-throughput sequencing of short DNA fragments that contain both the 5′UTR and the potentially flipped DNA sequence gives information on both 5′UTR genotype and related expression phenotype, which provided data on the recombinase expression from 300,000 different 5′UTRs, which were fed into machine learning. The analysis, surprisingly, revealed that, rather than mRNA secondary structures, the presence and positioning of the SD are most important for high protein production in this case, possibly because the 5′ end of the recombinase ORF was unlikely to form strong mRNA structures with any UTR. The machine-learning approach was used to develop a new 5′UTR design algorithm that Höllerer et al., 2020 report outperformed currently available algorithms, which are mostly based on biophysical models. However, this algorithm has not yet been tested for ORFs other than the recombinase ORF in that study.

Likewise, successful 5′UTR prediction algorithms based on multiple regression or machine learning approaches have been developed for yeast (Cuperus et al., 2017; Decoene et al., 2018; Ding et al., 2018). Such big-data analyses, based on randomized sequence libraries, seem a promising road toward better predictive algorithms for robust regulation of synthetic genes.

2.7 Conclusions

Despite significant efforts to elucidate the effect of codon usage and other gene features on protein production, it is still not completely understood. During the past decade, genome, transcriptome, proteome, and translatome (ribosome profiling) data became increasingly available. Bioinformatic analysis of those data has provided relevant insights into coding features and their relation to protein production. Recently, such analyses, combined with half-life measurements of mRNA, led to the discovery that optimal translation of an mRNA increases its stability in eukaryotes. However, many factors and their relevance are still unclear and require further investigation and, possibly, new experimental approaches.

One of the key knowledge gaps is the role of mRNA secondary structures, which is suggested to have a pivotal role in translation initiation and elongation, but its true effect is still unsettled. Recently, emerging protocols enabled the generation of transcriptome-wide *in vivo* mRNA structural data. However, groups using such methods report partly contradicting results for the role of mRNA secondary structures on translation-elongation efficiency (Burkhardt et al., 2017; Mustoe et al., 2018). Further refinement and validation of those protocols are required to improve the understanding of mRNA structures on translation. Another poorly explored territory is the influence of the ORF's codon sequence on co-translational folding and fidelity. Bioinformatic analysis of genome and translatome data suggested important roles for translation speed on protein folding, at least for some proteins. Detailed molecular studies focusing on some specific

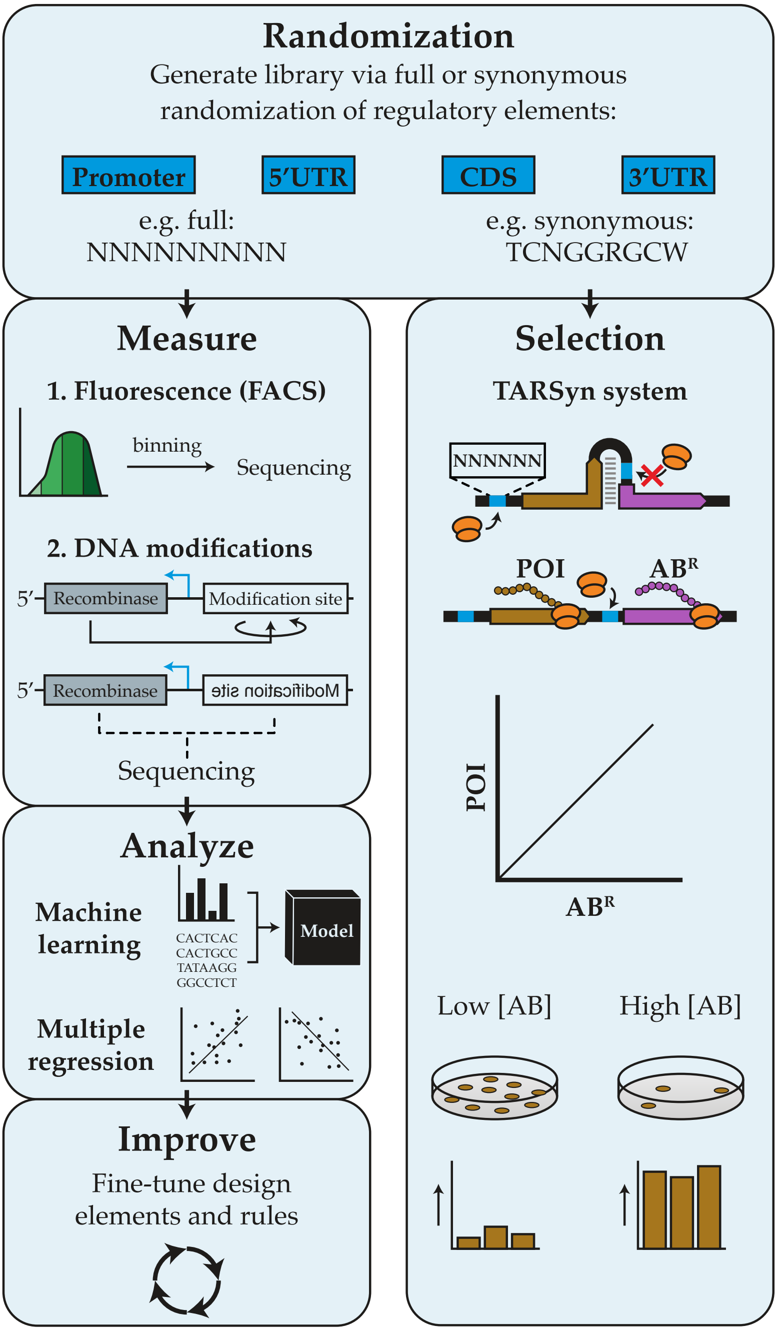

Randomization
Generate library via full or synonymous randomization of regulatory elements:
Promoter
5′UTR
CDS
3′UTR
e.g. full:
NNNNNNNNN
e.g. synonymous:
TCNGGRGCW
Measure
1. Fluorescence (FACS)
binning
Sequencing
2. DNA modifications
5′
Recombinase
Modification site
5′
Recombinase
Modification site
Sequencing
Selection
TARSyn system
NNNNNN
POI
AB^R
POI
AB^R
Low [AB]
High [AB]
Analyze
Machine learning
CACTCAC
CACTGCC
TATAAGG
GGCCTCT
Model
Multiple regression
Improve
Fine-tune design elements and rules

FIGURE 2.4: **Overview of a Typical Workflow Randomizing Gene Regulatory and ORF Sequences.** After randomization of genetic regulatory sequences or (part of) the codons of an ORF, the protein production by the resulting (large) variant library can be measured and binned according to fluorescence levels by fluorescent activated cell sorting (FACS). As an alternative to fluorescent-reporter proteins, a DNA-modifying enzyme can be used as a reporter because its expression can be assessed by high-throughput sequencing of modifications in the DNA. The latter approach was demonstrated for the expression of a randomized 5'UTR library mediating the expression of a recombinase that flips a nearby DNA modification site. In the same single-sequencing read, the 5'UTR variant can be identified and whether the site was flipped or not, allowing high-quality, large-scale data on expression levels (Höllerer et al., 2020). Analysis of generated large-scale data is typically performed by multiple regression analysis and, recently, by machine-learning algorithms. Next, understanding of expression levels can be further improved by correlations or rules derived from the analysis, and expression could be further studied during next-iteration rounds in which randomized sequence space can be limited, based on the results of the previous iterations. As an alternative to the learning cycle, a direct-selection system can be used for the selection of high-expressing variants. For example, the so-called TARSyn system allows for the selection of high-expressing clones based on antibiotic resistance (Rennig et al., 2018). The expression of a (non-reporter) protein of interest is translationally coupled to downstream antibiotic resistance, allowing for easy selection for high expression under high antibiotic concentrations.

proteins have confirmed that codon usage has a crucial role in folding. However, data and protocols to test this hypothesis experimentally for larger sets or proteins or on a proteome-wide scale are lacking.

A general limitation of studying genetic features within native genes (in a certain organism or under certain conditions) is the complexity in detecting "weak signals" from relevant factors within sequences that underwent optimization during millions of years of evolution. Alternative approaches, based on synthetic gene libraries, represent strong complementary methods in which many variants for a single gene can be generated to probe relevant factors. However, these "controlled" studies have, so far, been able to provide generic explanations for variable protein production levels only to some extent and are mostly based on correlating expression with known factors. In addition, those studies have mostly focused on a few highly expressed reporter proteins (mostly GFP), which may make conclusions biased.

Machine-learning approaches may help to further elucidate unknown features and factors in a more unbiased way. Such approaches have recently been applied to analyze expression data from randomized synthetic libraries of promoters and 5′UTRs. Such approaches may be promising for developing better predictive algorithms. However, large datasets are required for machine-learning algorithms to generate predictive models, and machine learning does not necessarily lead to increased biological understanding because, sometimes, such machine-learning approaches generate a predictive "black box."

The limited understanding of the fundamental rules in protein production remains a significant challenge for its applications. Problems in synthetic gene design are regularly observed for tuning and optimizing production of biotechnological or medical relevance. These challenges become even more pressing for synthetic biologists trying to construct designer genomes, which require tuning of many synthetic genes simultaneously.

Specific methods have been proposed that can, to some extent, increase the predictability of synthetic gene design. Typically, commercial or academic codon optimization algorithms are used to design ORF regions for heterologous expression, often with limited success, which is not surprising given the current knowledge gaps. However, promising design and randomization approaches have been established regarding the engineering of the highly influential region comprising the 5′UTRs and the first few codons of an ORF.

Overall, both for the understanding of the fundamental natural principles of gene design and expression and for diverse applications, there remains a need to delve further into the outstanding questions in this field. Despite the impressive recent progress, further refinement of recently launched techniques, as well as the development of new experimental and computational approaches, will be essential to address key questions that have intrigued many biologists for decades.

2.8 Acknowledgments and Author Contributions

T.N., M.F.-B., N.J.C., and J.v.d.O. acknowledge the support of the Netherlands Organisation for Scientific Research (NWO) via the Gravitation Project BaSyC (024.003.019) and Spinoza (SPI 93-537), awarded to J.v.d.O. In addition, N.J.C. acknowledges support from his NWO Veni fellowship (VI.Veni.192.156).

T.N., M.F-B., J.v.d.O., and N.J.C. reviewed the literature. T.N., M.F.-B., J.v.d.O, and N.J.C. wrote the review.

Chapter 3

Improved protein production and codon optimization analyses in *Escherichia coli* by bicistronic design

Adapted from:

Nieuwkoop, T., Claassens, N. J., & van der Oost, J. (2019). Improved protein production and codon optimization analyses in *Escherichia coli* by bicistronic design. *Microbial biotechnology*.

Abstract

Different codon optimization algorithms are available that aim at improving protein production by optimizing translation elongation. In these algorithms, it is generally not considered how the altered protein-coding sequence will affect the secondary structure of the corresponding RNA transcript, particularly not the effect on the 5'UTR structure and related ribosome binding site availability. This is a serious drawback, because the influence of codon usage on mRNA secondary structures, especially near the start of a gene, may strongly influence translation initiation. In this study, we aim to reduce the effect of codon usage on translation initiation by applying a bicistronic design (BCD) element. Protein production of several codon-optimized gene variants is tested in parallel for a BCD and a standard monocistronic design (MCD). We demonstrate that these distinct architectures can drastically change the relative performance of different codon optimization algorithms. We conclude that a BCD is indispensable in future studies that aim to reveal the impact of codon optimization and codon usage correlations. Furthermore, irrespective of the algorithm used, using a BCD does improve protein production compared with an MCD. The overall highest expression from BCDs for both GFP and RFP is at least twofold higher than the highest levels found for the MCDs, while for codon variants having very low expression from the MCD, even 10-fold to 100-fold increases in expression were achieved by the BCD. This shows the great potential of the BCD element for recombinant protein production.

3.1 Introduction

Heterologous protein production in prokaryotes is one of the major hallmarks of biotechnology and synthetic biology, and it forms the foundation of a wide range of medical and industrial innovations (Elena et al., 2014). However, optimization of protein production mostly relies on a trial-and-error approach. The poor predictability of high-level protein production is due to the complexity and interconnection of several determining factors. Key factors at the transcriptional level are the gene's copy number and promoter strength. At the translational level, the ribosome binding site (RBS) strength, mRNA secondary structure and codon usage are key factors that together play a major role in efficient protein production (Kudla et al., 2009; Mutalik et al., 2013b; Rosano and Ceccarelli, 2014; Quax et al., 2015). Especially, factors at the translational level are highly complex, and our limited understanding of these interconnected factors often hampers high protein production (Mutalik et al., 2013b; Quax et al., 2015).

Translation initiation in prokaryotes occurs when the 16S rRNA of the small ribosomal subunit binds the RBS in the 5'UTR of a gene. After this, the large ribosomal subunit is recruited and translation elongation can start. The RBS must be freely accessible to allow recruitment of the ribosomal subunits. Hence, strong secondary structures in the mRNA involving the RBS result in poor ribosome binding kinetics (Studer and Joseph, 2006), which can lead to reduced protein production (Smit and Duin, 1990; Kudla et al., 2009; Salis, Mirsky, and Voigt, 2009; Goodman, Church, and Kosuri, 2013). Secondary structures that include the RBS motif have been reported to form either via local contacts between the 5'UTR and the adjacent start of the coding sequence (CDS) or via long-range interactions through base pairing of the 5'UTR with more distal regions in the CDS (Mustoe et al., 2018). A constant 5'UTR region can, therefore, perform differently regarding translation efficiency in the case of different CDS and 3'UTR sequences (Griswold et al., 2003). In extreme cases, secondary structures between the RBS and CDS have been reported to block translation completely (Mutalik et al., 2013a; Mirzadeh et al., 2015).

Given the degeneracy of the genetic code, 61 codons for only 20 amino acids, many different codon sequence variants can encode a certain protein. During translation elongation, codon usage is a crucial factor that can influence the efficiency of protein production in multiple ways. The elongation rate can be limited by several factors such as the availability of cognate aminoacyl-tRNA's (Hanson and Coller, 2018) and the presence of potential hurdles in the CDS, such as RBS-like sequences (Li, Oh, and Weissman, 2012; Vasquez et al., 2016) and secondary structures (Takyar, Hickerson, and Noller, 2005; Buchan and Stansfield, 2007; Chen et al., 2013). Coding sequences that are efficiently translated were also reported to be linked to longer mRNA lifetimes, further enhancing production (Boël et al., 2016). Whereas in native situations, codon usage has been extensively tuned in the course of evolution, attempts to express such genes at very high levels in heterologous production hosts are often hampered. This can potentially be solved by substituting the codons with synonymous counterparts. However, transcript secondary structure and codon sequence are intrinsically correlated. Therefore, the effect of single or multiple synonymous codon substitutions cannot be clearly attributed to changes in translation elongation or in translation initiation (Gustafsson et al., 2012; Gorochowski et al., 2015).

Many codon optimization algorithms have been developed aiming to improve heterologous protein production (Gould, Hendy, and Papamichail, 2014), although with varying success rates in terms of increased functional protein production (Maertens et al., 2010; Gustafsson et al., 2012; Claassens et al., 2017; Mignon et al., 2018). This variety can be partially explained by the introduction of new secondary structures within the transcript due to synonymous codon changes (Nørholm et al., 2013; Mirzadeh et al., 2015). Particularly, secondary structures at the 5′UTR are overlooked as most optimization algorithms only consider optimization of the CDS and do not take the 5′UTR into account. Still, when the 5′UTR sequence would be included in the design, currently available tools for RNA secondary structure prediction are not accurate enough to robustly design well-accessible 5′UTRs.

To properly study the effects of codon usage and codon optimization approaches on translation elongation, the effects of codons on translation initiation need to be decoupled. To some degree, secondary structures at the 5′UTR can be predicted *in silico*, and synonymous codons can be introduced to remove these limitations. However, this requires custom design for each construct and limits codon studies as it dictates codons at the start of the gene. Alternatively, the undesired 5′UTR structure may be solved, either on purpose or accidentally, by including well-expressed N-terminal protein fusions in the expression vector. These fusions are mostly included to facilitate affinity purification or folding for specific proteins (e.g. His-tag or MPB-tag; Griswold et al., 2003; Vasquez et al., 2016). However, the addition of an N-terminal peptide to the protein may affect protein functionality and may require additional cleavage and hence is not always a desirable solution.

Therefore, we decided to use a bicistronic design (BCD) element controlling expression of heterologous genes (Makoff and Smallwood, 1990). These elements were previously developed by Mutalik et al., 2013a for reliable generic control of different genes. The BCD contains a well-accessible RBS1 motif that drives the translation of a short peptide (Figure 3.1a). Within the short peptide's CDS, RBS2 is present that allows for translation initiation of the protein of interest, and the stop codon of the peptide sequence overlaps with the start codon of the target CDS. This genetic architecture leads to the translational coupling of the short peptide to the protein of interest (Mutalik et al., 2013a). After transcription of the bicistronic mRNA, the ribosome readily binds to the well-accessible RBS1 site and translates the first cistron; then, the RBS2 site probably becomes available due to the intrinsic helicase activity of the ribosome, irrespective of adverse mRNA secondary structures (Takyar, Hickerson, and Noller, 2005).

We here describe the effects of a BCD element on the expression of various codon-optimized variants of the green fluorescent protein from *Aequorea victoria* jellyfish, optimized for excitation by UV light (GFPuv; Crameri et al., 1996), and a monomeric version of the red fluorescent protein from *Discosoma* coral (mRFP; Campbell et al., 2002). Both proteins are of eukaryotic origin, which makes them good models for studying codon optimization in a distant bacterial expression host, while their functional expression levels can be easily estimated by measuring fluorescence.

Production from BCDs is compared with production as a single gene (monocistronic design, MCD), the architecture that is generally used for heterologous protein production. We demonstrate that these BCD elements can positively influence the performance of different codon optimization algorithms. Hence, we propose that these BCD elements should be an essential part of future codon usage studies to eliminate the potentially overlapping influence of RNA secondary structure.

3.2 Results and discussion

Various optimized coding sequences for mRFP and GFPuv were expressed using the relatively weak, constitutive beta-lactamase promoter (Pbla). The low transcription rate prevents possible oversaturated gene expression and as such generates a dynamic range that allows for accurately comparing the effects of the used codon optimization strategies and of the BCD and MCD elements. The regularly used (RU) mRFP (Campbell et al., 2002) and GFPuv (Crameri et al., 1996) sequences, both containing several distinctive mutations compared with the wild type for better stability of fluorescence properties, were compared with several other codon variants, all having identical amino acid sequences to the regularly used protein. These variants include a codon-harmonized (H) variant (Angov, Legler, and Mease, 2011), a multiparameter codon-optimized variant generated using GeneArt's GeneOptimizer software (Opt; Raab et al., 2010) and a tRNA codon-optimized (tRNA) variant (Table S3.2). Codon harmonization copies the codon usage landscape from the original host to the new host (Angov, Legler, and Mease, 2011; Claassens et al., 2017). GeneArt's GeneOptimizer algorithm performs multiparametric optimization with an apparent preference for common codons as it generated a sequence with the highest Codon Adaptation Index (> 0.9, Table S3.2; Raab et al., 2010). The tRNA codon optimization replaces codons for codons that have the highest number of complementary tRNA genes. Additionally, a transcript was designed with minimal overall mRNA secondary structure including the fixed 5'UTR and 3'UTR regions (codon usage variants based on this will hereafter be referred to as dG), which allows all possible codons. Lastly, a minimal overall free folding energy transcript was included, which is restricted to codons with well-represented tRNA's (tRNA-dG). The harmonized sequence for mRFP is not included, as it could not be designed because the genome of its original host, *Discosoma sp.*, is not available.

Protein production overall increases when using a BCD compared with MCD for all GFPuv variants (Figure 3.1b). The harmonized and optimized GFPuv sequences resulted in increased protein production compared with the RU sequence in combination with the MCD 5'UTR. The tRNA, tRNA-dG and dG variants with an MCD 5'UTR led to lower protein production versus the RU gene variant. However, when comparing protein production of variants expressed with a BCD, completely different relative expression ratios are observed. The harmonized variant performed worse than the RU sequence, and the expression of the tRNA-optimized sequence was similar to that of the RU sequence. The two transcript variants designed to have a low overall free energy (tRNA-dG and dG) had reduced expression compared with the RU sequence; however, the addition of the BCD improved expression for both variants versus the MCD.

For the mRFP expression, similar effects of the BCD were observed. The overall mRFP production improved by the BCD, and relative differences among codon variants are very different compared with the MCD (Figure 3.1c). As an exception, the expression of the *E. coli* optimized mRFP did not benefit from the BCD but stayed equal, suggesting that translation initiation is not the limiting factor in this case.

Although the specific codon optimization methods applied in this study were not the main focus, some conclusions can be drawn regarding these methods. First, there is no algorithm that consistently stands out for optimal production of both GFPuv and mRFP. Secondly, a decrease in transcript free energy, especially

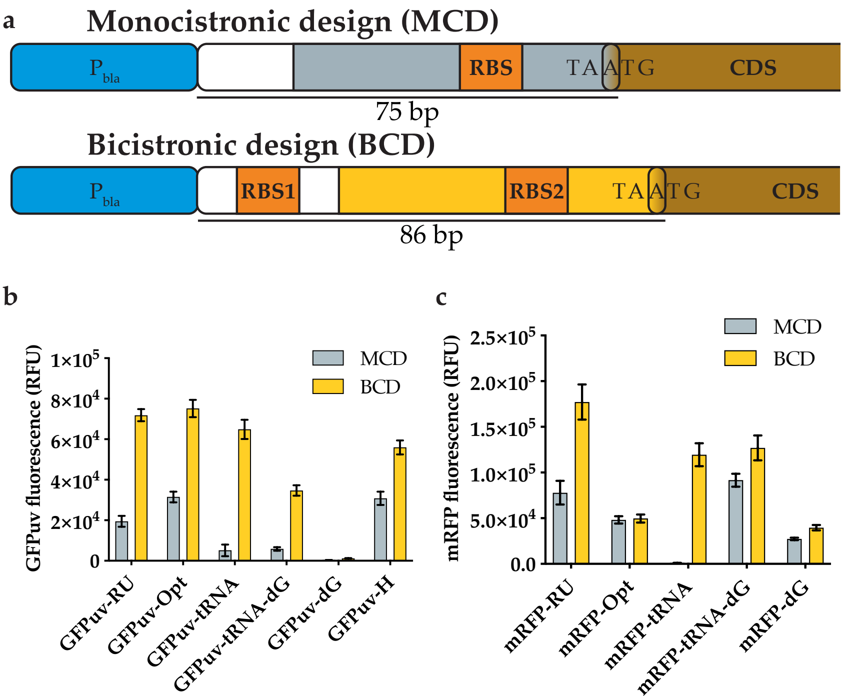


FIGURE 3.1: **The effects of a BCD on gene expression. (a)** Genetic architecture of monocistronic and bicistronic design. **(b, c)** The effect of a bicistronic and monocistronic design on the expression of different codon-optimized GFPuv **(b)** and mRFP **(c)** variants (RFU: relative fluorescence units). The regularly used GFPuv (GFPuv-RU) sequence is compared to an optimized sequence (GFPuv-Opt), *Escherichia coli* tRNA-optimized sequence (GFPuv-tRNA), *E. coli* tRNA-optimized sequence with subsequent minimalized free energy (GFPuv-tRNA-dG), a minimal free energy transcript (GFPuv-dG) and an *E. coli* harmonized sequence (GFPuv-H). The regularly used mRFP sequence (mRFP-RU) is compared with the *E. coli* optimized sequence (mRFP-Opt), *E. coli* tRNA-optimized sequence (mRFP-tRNA), *E. coli* tRNA-optimized sequence with subsequent minimalized free energy (mRFP-tRNA-dG) and a minimal free energy transcript (mRFP-dG). Production is determined using flow cytometry for eight biological replicates for each variant. The error bars depict the standard deviation for the average expression of eight biological replicates. For each replicate, the expression level of 50 000 single cells is measured, averaged and normalized to a cell culture not expressing any fluorescent protein. For all cases, except mRFP-Opt, the fluorescence of the BCD variants over the MCD variants is significantly different at a P-value of 0.001. Similar results are obtained for fluorescence measurements obtained with a plate reader (Figure S3.4). The MCD and BCD sequence can be found in Table S3.1.

for the dG variants, seems to lead to reduced expression, possibly due to the incorporation of rare codons in favour of low secondary structures (CAI score < 0.55, Table S3.2).

In the case of the mRFP-tRNA variant, we further investigated the surprisingly large increase in production from the BCD relative to the MCD (over 100-fold). The extremely low production from the MCD might be explained by a seriously hampered RBS accessibility. In this specific case, *in silico* secondary structure analysis of the MCD mRFP-tRNA transcript indeed revealed that the RBS site was involved in a strong loop (Figure 3.2a), which could prevent the ribosome from binding. This structure is also predicted in the BCD construct (Figure 3.2b); however, the BCD expression appeared not to be affected, as was expected based on the functionality of the BCD architecture that generally prevents issues with RBS2 inaccessibility, probably through the aforementioned ribosome helicase activity (Takyar, Hickerson, and Noller, 2005; Mutalik et al., 2013a). With an *in silico* prediction; we attempted a design to weaken the RBS-containing secondary structure by introducing a silent point mutation in the CDS (Figure 3.2a and c). Experimentally, it could indeed be demonstrated that this mutation recovered expression of mRFP-tRNA with the MCD, at levels similar to those of the BCD (Figure 3.2d).

While the translation initiation limitation for mRFP-tRNA could be obviously predicted using *in silico* mRNA structure analysis, this was not that obvious for the other expressed GFPuv (Figure S3.1) and mRFP (Figure S3.2) constructs. Likewise, expression levels for MCD constructs did not correlate with predictions by the RBS Calculator algorithm (Salis, Mirsky, and Voigt, 2009; Espah Borujeni, Channarasappa, and Salis, 2014; Figure S3.3). This again shows the general limitation of biophysical models and *in silico* tools to design reliable UTR's, whereas the BCD system does not depend on such tools.

Our results show the importance of an accessible RBS region for overall translation efficiency. Due to the intrinsic correlation between the coding sequence and secondary structures of the corresponding mRNA, it will be hard to disentangle these factors in correlation studies. Further, we note that the overall increased expression may also be partly caused by a higher number of ribosomes sequestered to translate the ORF due to the presence of two RBSs. Generally, using a BCD may eliminate the translation initiation as the rate-limiting step of the translation process. Hence, the BCD approach seems the way to go to study the effect of synonymous codon substitutions on protein production in *E. coli*, and likely also in other prokaryotes. For potential issues related to translation initiation in

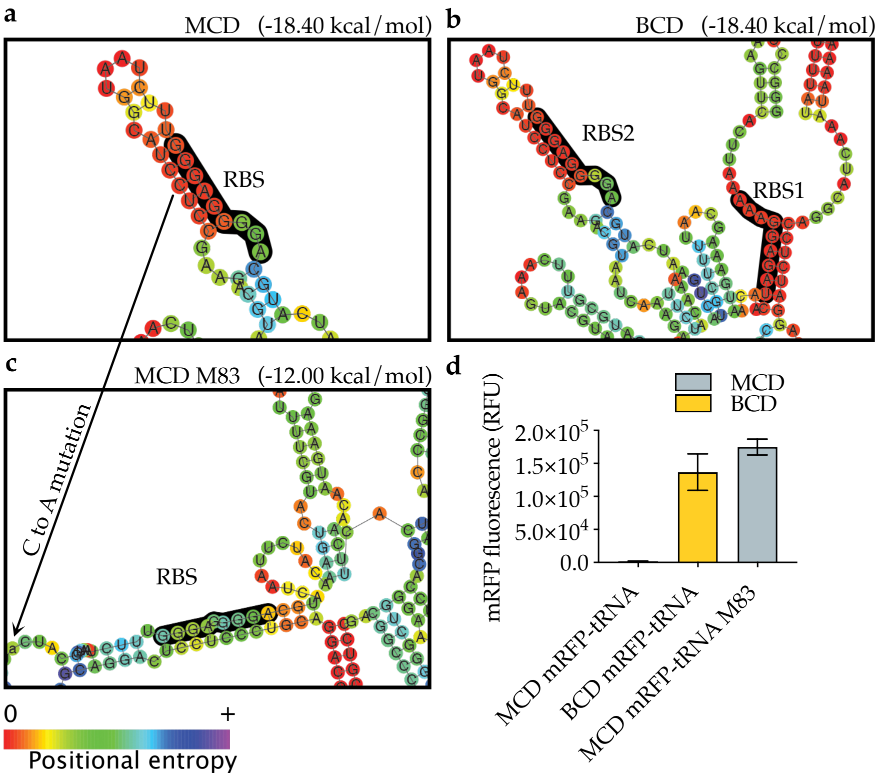


FIGURE 3.2: **Secondary structure predictions and effects on expression.** **(a)** Secondary structure prediction of the mRFP-tRNA transcript with a monocistronic design. The arrow indicates the nucleotide that was silently mutated in an attempt to dissolve the structure (M83; C → A. 5′ and 3′ indicate the orientation of the RBS). **(b)** Secondary structure prediction of the mRFP-tRNA transcript with a bicistronic design. **(c)** Secondary structure prediction of the mRFP-tRNA M83 transcript with a monocistronic design. The RBS sites are highlighted in black, and the positional entropy for each nucleotide is indicated with a colour gradient. The free energy for each construct is calculated with a sequence window containing the 5′UTR and the first 36 nucleotides of the CDS. **(d)** Relative mRFP expression of the mRFP-tRNA with the MCD, BCD and MCD M83 mutation (RFU: relative fluorescence units). The error bars depict the standard deviation for the average expression of eight biological replicates. For each replicate, the expression level of 50 000 single cells is measured, averaged and normalized to a cell line without mRFP. The mean differences are significantly different at a P-value of 0.001.

eukaryotes, different tools will be required, as they rely on fundamentally different translation initiation mechanisms. However, previously developed tools based on upstream open reading frames (uORFP) may be a useful eukaryotic tool (Morris and Geballe, 2000; Ferreira, Overton, and Wang, 2013), somewhat analogous to BCDs in prokaryotes. Finally, it is concluded that the outcome of the here used codon optimization methods is still rather unpredictable, and better, consistently performing codon optimization algorithms need to be explored, such as by Design of Experiment approaches (Gustafsson et al., 2012). An interesting outcome of this study is that the experimental data do confirm the promise of using BCD elements as a generic approach to increase yields in heterologous protein production (Roy et al., 2017).

3.3 Materials and methods

3.3.1 Strains and media

E. coli DH10B (Invitrogen) was used as cloning and expression host. *E. coli* was cultivated in LB (Lysogeny Broth) and LB agar with 50 µg/mL kanamycin when appropriate.

3.3.2 Strain and plasmid construction

Plasmid pFAB3909 (Addgene #47812) was used as an expression vector. pFAB3909 is designed by Mutalik et al., 2013a and contains a bicistronic design (BCD). The BCD5 variant was selected for its relatively high RBS strength and low variance in production for different genes. The promoter sequence was replaced with a bla promoter via PCR using NEB's Q5® High-Fidelity DNA Polymerase according to the standard protocol. An oligo with the bla promoter as overhang (Oligo 10023, Table S3.3) and phosphorylated oligo (Oligo 10024) bind on either side of the P14 promoter. The resulting amplification fragment is cleaned and concentrated using Zymo's DNA Clean & Concentrator kit according to the standard protocol. The cleaned-up DNA is ligated using NEB's T4 DNA Ligase overnight at room temperature. The ligated DNA is transformed into *E. coli* DH10B using heat shock and the sequence is confirmed (Macrogen). The MCD expression vector was constructed by deleting the RBS1 using the same method as described previously (Oligo 10805 and 10806). To introduce the mutation, to abolish the secondary structure around the RBS in pTN0004_mRFP-tRNA, a silent point mutation was introduced using the same method as described previously (Oligo 11665 and 11666). The mRFP(Campbell et al., 2002) and GFPuv sequences were ordered as gBlocks (IDT) and seamlessly cloned into the pFAB3909 vector using NEBuilder® HiFi DNA Assembly Master Mix.

3.3.3 Protein quantification by flow cytometry

Eight single colonies were picked from an overnight agar plate, containing transformants, and each used to inoculate 200 µl medium in a 2 mL 96 wells plate (Greiner Bio-One, V-bottom). The cultures were grown at 37°C for 18 hours at 200 rpm (1-inch stroke). The cultures were diluted 1000 times in 1x PBS and measured using the Attune NxT flow cytometer (ThermoScientific, software version 2.5). 50,000 single cell events were used to obtain the average mRFP fluorescence (excitation 561 nm, emission 620/15 nm) or GFPuv fluorescence (excitation 405 nm, emission 512/25 nm) for each biological replicate. The mean fluorescence of each replicated was corrected by subtracting the average fluorescence of an *E.coli* strain not expressing mRFP or GFPuv.

3.3.4 Protein quantification by plate reader

Eight colonies were picked from an overnight agar plate, containing transformants, and each used to inoculate 300 µl medium in a 2 mL 96 wells plate (Greiner Bio-One, V-bottom). The cultures were grown at 37°C for 18 hours at 200 rpm (1-inch stroke). 200 µL culture was transferred to a 200 µl 96 wells plate (V-bottom) and centrifugated for 10 minutes at 3800g. The pellets were washed twice with 200 µL 50 mM Tris HCl pH 7.5. 100 µL was transferred to a black sided clear bottom 96 wells plate and the fluorescence was measured using a Synergy Mx plate reader (BIOTEK, software version 3.02.1) (mRFP excitation at 586/9 nm, emission 661/9 nm, gain 125 and GFPuv excitation at 399/9 nm, emission 510/9 nm, gain 75). The fluorescence of each biological replicate was normalized using the OD_{600} and corrected by subtracting the fluorescent value of an *E.coli* strain not expressing mRFP or GFPuv.

3.3.5 Codon optimization algorithms

Codon harmonization (H) of the mRFP and GFPuv CDS has been performed using our online Codon Harmonizer tool (codonharmonizer.systemsbiology.nl) (Claassens et al., 2017) based on the original algorithm by Angov et al., 2008. Codon optimization has been performed using GeneArt's GeneOptimizer algorithm web tool (Raab et al., 2010) (performed in May 2017). tRNA optimization was performed in-house by replacing all codons with codons that are represented by a tRNA with the highest genome copy number with a preference of Watson-Crick base pairing over wobble base pairing. Genomic tRNA copy numbers for *E. coli* DH10B were derived from gtrnadb.ucsc.edu. dG optimization consisted of random synonymous mutations to lower the overall minimal free folding energy of the construct as much as possible without codon limitations. An in-house script was developed for this purpose (Available in online publication dx.doi.org/10.1111/1751-7915.13332). tRNA-dG optimization is a combination of the previously mentioned optimization methods. The lowest overall minimal free energy is desired while only codons are selected that are well-represented by tRNAs.

3.3.6 mRNA secondary structure analysis

The ViennaRNA Package (Lorenz et al., 2011) was used for all mRNA secondary structure predictions. The RNAfold program (version 2.4.3) was used to generate the minimum free energy secondary structure and base pairing probability matrix. The RNAplot program (version 2.4.3) and Relplot.pl algorithm (version 1.3) was used to draw the RNA secondary structures with base pair probability and highlighted RBS sites.

3.4 Acknowledgement

We would like to thank Sjoerd Creutzburg for providing the harmonized GFPuv gene. TN and JvdO are supported by The Netherlands Organization for Scientific Research (NWO) through a Gravitation Grant (Project 024.003.019). NJC is supported by NWO through a Rubicon Grant (Project 019.163LW.035).

3.5 Supplementary information

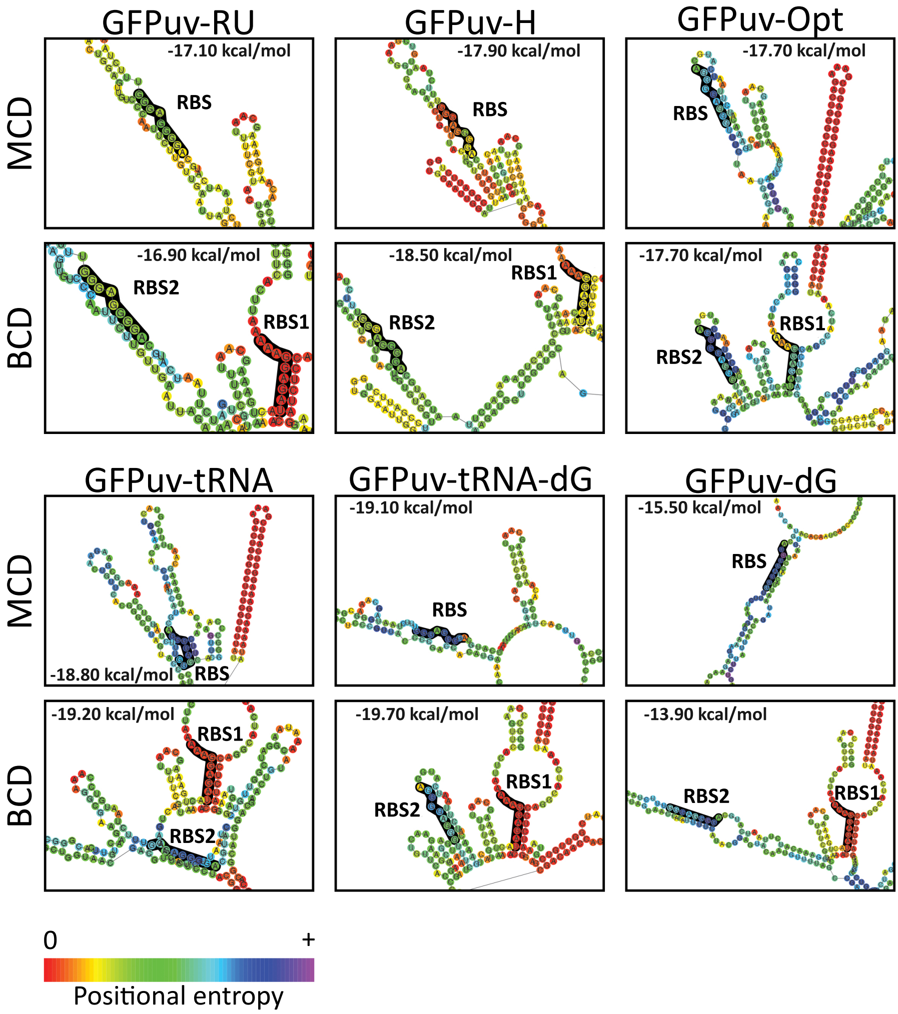


FIGURE S3.1: **5′UTR secondary structure predictions for differently codon optimized GFPuv transcripts.** Either with an MCD element or a BCD element. Ribosome binding sites (RBS) are indicated in black and the positional entropy for each nucleotide is indicated with a color gradient. The free energy for each construct is calculated with a sequence window containing the 5′UTR and the first 36 nucleotides of the CDS.

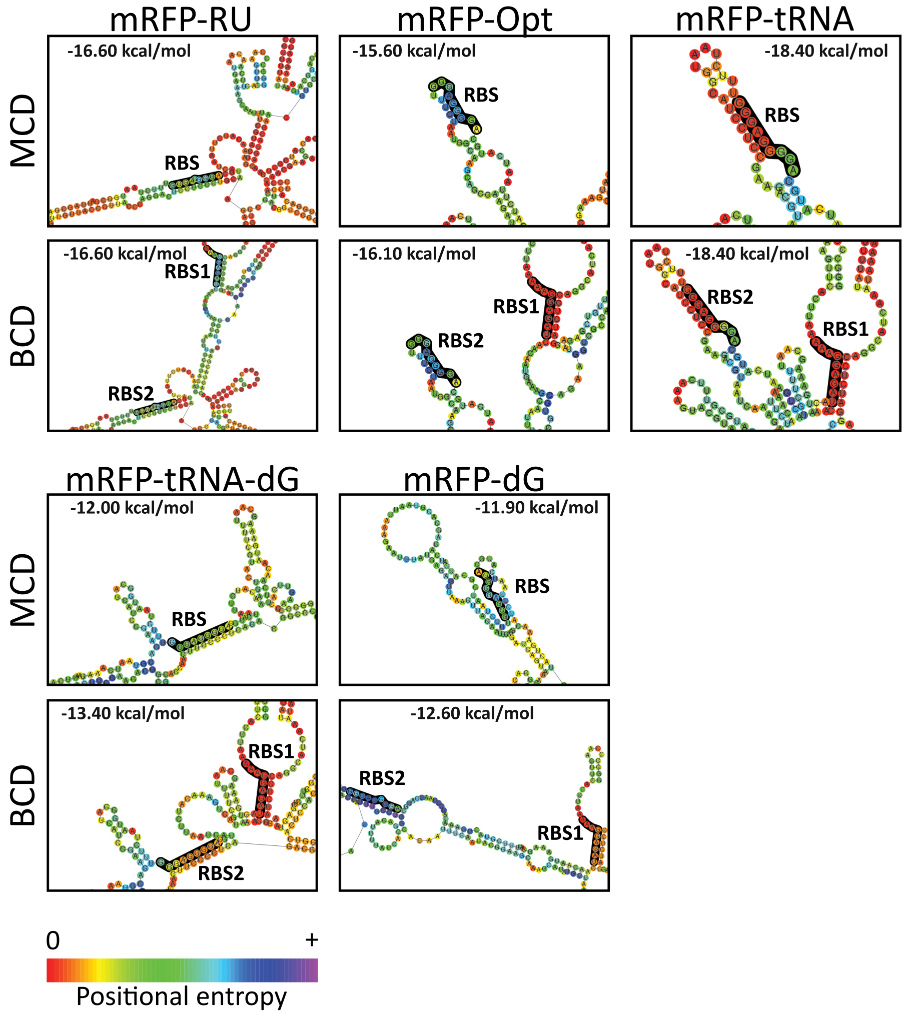


FIGURE S3.2: **5′UTR secondary structure predictions for differently codon optimized mRFP transcripts.** Either with an MCD element or a BCD element. Ribosome binding sites (RBS) are indicated in black and the positional entropy for each nucleotide is indicated with a color gradient. The free energy for each construct is calculated with a sequence window containing the 5′UTR and the first 36 nucleotides of the CDS.

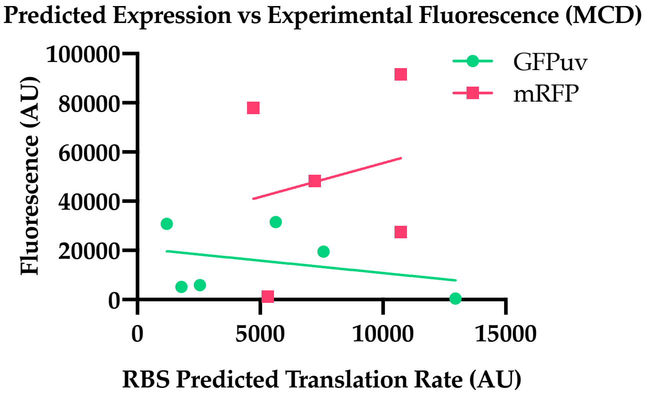


FIGURE S3.3: **Comparison between predicted translation rate by RBS calculator and experimentally determined expression of protein for MCD constructs.** There is no correlation for mRFP (p-value of 0.7275) or GFPuv (p-value of 0.5219).

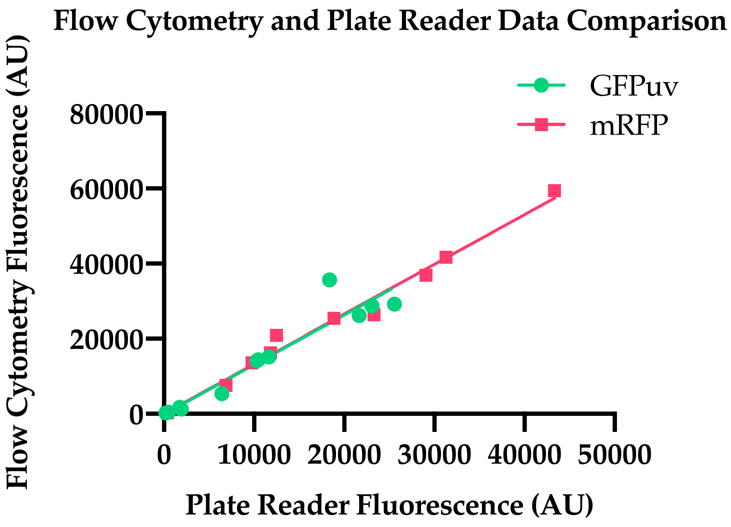


FIGURE S3.4: **Correlation between single-cell data obtained with flow cytometry and OD600 corrected bulk fluorescence using a plate reader.** Both the measurement for GFPuv (R^2CV = 0.824344) and mRFP (R^2CV = 0.9752014) correlate well.

TABLE S3.1: **DNA sequences of different genetic elements used.**

Region	**DNA sequence**
bla promoter	1 TTCAAATATG TATCCGCTCA TGAGACAAT
BCD	1 GGGCCCAAGT TCACTTAAAA AGGAGATCAA 31 CAATGAAAGC AATTTCGTAC TGAAACATCT 61 TAATCATGCA GGGGAGGGTT TCTAA
MCD	1 GGGCCCAAGT TCACTTCAAC AATGAAAGCA 31 ATTTTCGTAC TGAAACATCT TAATCATGCA 61 GGGGAGGGTT TCTAA
RBS1	1 AAAGGAGAT
RBS2	1 AGGGGAGGG
spacer + rrnB1 terminator	1 GGATCGGTTG TCGAGTAAGG ATCTCCAGGC 31 ATCAAATAAA ACGAAAGGCT CAGTCGAAAG 61 ACTGGGCCTT TCGTTTTAT

TABLE S3.2: **CDS codon adaptation index (CAI) and Gibbs free energy calculation of both the MCD and BCD variant.** For free energy calculations the whole 5′UTR region and the first 12 codons of the CDS are used. The complete DNA sequence of each CDS variant can be found in the supporting information of the online publication (dx.doi.org/10.1111/1751-7915.13332).

CDS variation	**CAI**	**ΔG MCD(kcal/mol)**	**ΔG BCD(kcal/mol)**
GFPuv-RU	0.61	-17.10	-16.90
GFPuv-Opt	0.91	-17.70	17.70
GFPuv-tRNA	0.79	-18.80	-19.20
GFPuv-tRNA-dG	0.78	-19.10	-19.70
GFPuv-dG	0.55	-15.50	-13.90
GFPuv-H	0.70	-17.90	-18.50
mRFP-RU	0.82	-16.60	-16.60
mRFP-Opt	0.90	-15.60	-16.10
mRFP-tRNA	0.80	-18.40	-18.40
mRFP-tRNA-dG	0.78	-12.00	-13.40
mRFP-dG	0.54	-11.90	-12.60
mRFP-tRNA M83	0.80	-12.00	-13.40

TABLE S3.3: **Oligo sequences** (P indicates a phosphorylated 5′-end).

Oligo ID	**Sequence (5′ - 3′**
10023	1 TTCAAATATG TATCCGCTCA TGAGACAATG 31 GGCCCAAGTT CACTTAAAAA GG
10024 (P)	1 GTTATGCAGC AACGACTCAT AGAAAG
10805	1 CAACAATGAA AGCAATTTTC GTAC
10506 (P)	1 AAGTGAACTT GGGCCC
11665 (P)	1 CGAAGACGTA ATCAAAGAAT TC
11666	1 GATGATGCCA TTAGAAACCC

Chapter 4

Finding determinants of protein translation efficiency via codon randomization and machine learning

Thijs Nieuwkoop*, Barbara Terlouw*, John van der Oost, Nico Claassens

Laboratory of Microbiology, Wageningen University,
Stippeneng 4, 6708 WE Wageningen, The Netherlands

Abstract

Current codon optimization methods that aim for elevated protein production appear to be limited in their success. To elucidate which features are determinants of gene expression, and to identify rate-limiting sequence regions, we generated a library of the gene encoding a Red Fluorescent Protein (RFP) in *E. coli*, in which synonymous codons were randomized throughout the entire coding sequence (CDS). Using this library, we selected 1459 variants for analyzing complex correlations between the gene's nucleotide sequence and protein's functional production level (fluorescence). A wide range of expression levels was observed, of which the highest expressing gene variants, despite containing rare codons, outcompeted sequences generated through common codon optimization algorithms. By applying machine learning, we show that the bases surrounding the start codon and ribosome binding site are more important determinants for RFP protein production levels than codon usage throughout the gene. Specifically, the major predictive power ($r = 0.803$) concerns the identity of the 2nd to the 9th codon. The latter conclusion, based on randomization of the whole *rfp* CDS, is in agreement with that of earlier studies that relied on the analysis of more restricted *gfp* gene fragments. Assuming that our findings on the *rfp* gene reflect a general phenomenon, this would imply that codon optimization algorithms should be adapted accordingly.

4.1 Introduction

Due to degeneracy in the genetic code, a protein can be encoded by multiple codon sequences. While different synonymous codons do not alter the amino acid sequence, they are well-known to influence translation efficiency and co-translational protein folding (Buhr et al., 2016; Faure et al., 2016; Faure et al., 2017; Kim et al., 2015; Nieuwkoop et al., 2020; Zhang, Hubalewska, and Ignatova, 2009). Many codon optimization strategies have been developed that try to predict the most optimal codon usage to achieve high levels of protein production (Gould, Hendy, and Papamichail, 2014; Ranaghan et al., 2021). However, these algorithms often lack experimental validation where a sufficient number of sequences is tested. Optimization strategies vary widely and an experimental comparison study for these algorithms on a large set of proteins has not yet been performed (Ranaghan et al., 2021). The general idea behind most optimization algorithms is to design sequences such that readily available tRNAs are used as much as possible during translation to achieve optimal efficiency. In most optimization algorithms, the "optimality" of a codon is scored based on the occurrence of codons within coding regions (CDSs) of the expressing organism, either throughout the whole genome or in a subset of highly expressed genes. One measure of codon optimality is the Codon Adaptation Index (CAI): the geometric mean of the relative codon usage within a CDS. A CDS with a high CAI primarily uses frequent codons, while a CDS with a low CAI contains more rare codons. Although the CAI has become the most widely used metric in codon optimization algorithms, several studies report that no correlation exists between the CAI and protein production levels (see below) (Kudla et al., 2009; Welch et al., 2009). Therefore, using the CAI as an indicator for protein production levels may be an oversimplification, and many other contributing factors that are also related to codon usage may be more relevant (Nieuwkoop et al., 2020). This appears to be in agreement with the daily practice of many researchers, as obtaining sufficient protein production still regularly fails (Parret, Besir, and Meijers, 2016) and it still seems like a matter of trial-and-error.

In order to gain more insight into codon usage effects, Kudla *et al.* codon randomized the full CDS of the *gfp* gene, encoding the Green Fluorescent Protein (GFP) originating from *Aequoera victoria* (Kudla et al., 2009). Out of the 240 codons, 226 were synonymously mutated. Codon variants were generated from degenerate oligos using overlap extension PCR, yielding a total of 154 variants. While they did not find any correlation between expression levels and the CAI or the frequency of optimal codons, there was a correlation between expression and mRNA secondary structures in and near the 5′UTR. Interestingly, they did see a correlation between the CAI and the growth rate of the expression host, likely due to ribosome sequestration on transcripts containing rare codons (Andersson and Kurland, 1990). To explain that, Kudla *et al.* argued that selection for high CAI becomes relevant for highly expressed genes as a means to prevent the trapping of ribosomes and thereby facilitating the continued translation of essential genes. Other studies that focused on analyzing codon usage effects did limit their randomization to smaller sections of the gene, mainly the 5′ end of the CDS and the 5′UTR (Cambray, Guimaraes, and Arkin, 2018; Goodman, Church, and Kosuri, 2013; Frumkin et al., 2017). Here too, the general conclusion was that secondary structures including the 5′UTR are the main determinants for translation efficiency, alongside initial slow translation elongation rates due to the presence of so-called "translational ramps" (Frumkin et al., 2017; Tuller et al., 2010). The common theme in all these studies is that they used variations of the same reporter gene (GFP).

In order to study and discover factors that influence the overall level of protein production, we opted for full codon randomization of the gene encoding mRFP, originating from an anemone-like marine animal (*Discosoma*). We developed an assembly method to generate a library of synonymous codon variation along the entire CDS. We consider randomization of the complete CDS an absolute requirement for in-depth analysis of the implications of codon variation. A downside of full randomization is that it quickly leads to an inconceivably large number of possible variations. The *rfp* gene, which was randomized in this study by combining synonymous codons at all positions, includes 3.19e104 unique sequences, all encoding mRFPs with the same amino acid sequence. Clearly, only a minuscule fraction of the complete sequence space can be practically covered in such a randomized library. However, we wanted to test if with a small but high-quality data set from such a library we could harness the power of machine learning to elucidate important determinants in the coding sequence.

As also noted previously, mRNA secondary structures near the 5′UTR can greatly influence the translation initiation efficiency and therefore limit the overall translation process. Surprisingly, most popular codon optimization algorithms do not require including the 5′UTR sequence, while synonymous codon changes introduced by such algorithms have the potential to form or disrupt secondary structures within the whole mRNA. Removal of inhibitory structures is complicated by the limited *in silico* predictability of mRNA secondary structures. The length of mRNA molecules limits the reliability of secondary structure predictions, particularly long-distance interactions, as opposed to the better predictable structures of shorter molecules such as tRNA and miRNAs (Lange et al., 2012). Due to the limited reliability of mRNA folding predictions, it is difficult to attribute the effects of codon composition to either changes in the efficiency of translation elongation or to changes in mRNA secondary structure. To better study these determinants of gene expression separately, they need to be decoupled first. In this study we use a previously established genetic element called a bicistronic design (BCD) (Mutalik et al., 2013a, see Chapter 3), aiming to limit the influence of secondary structures in the 5'UTR as a major effect on overall protein production to allow us to study the more nuanced features associated with codon usage (Nieuwkoop, Claassens, and Oost, 2019). To complement the focus on the more nuanced features, we also used the relatively weak native *bla* promoter. A weak promoter will pose less strain on the transcriptional and/or translational processes, preserving the full expression range.

Furthermore, in this study, we opted for fairly low-throughput data acquisition by employing Sanger sequencing and individual culture measurements instead of single-molecule sequencing and FACS (Flow-seq). Flow-seq typically employs short-read sequencing (Illumina), which will not cover the full CDS length in a single read and due to the codon degeneracy, it would be difficult to assemble reads into contigs. Alternative long-read single-molecule methods (e.g., PacBio) would offer a solution. However, it was questionable whether a high enough coverage could be achieved to reach meaningful conclusions with such an approach (Peterman and Levine, 2016). The natural fluorescence variability between cells of the same culture is very large, increasing the likelihood that individual cells are binned wrongly and that the resulting dataset is too noisy to be analysed meaningfully through statistical analyses and machine learning.

We believe that machine learning particularly will be an incredibly valuable tool in solving the complex problem of codon optimization. When sufficient high-quality data is available, machine learning algorithms can discover complex correlations in raw data and annotated features that may be incomprehensible to the human eye. Additionally, some algorithms can be reverse engineered to reveal which features contribute to the predictability of the algorithm. Concretely, this means that machine learning not only can help us predict expression levels and protein activity, it also can help us understand the mechanisms behind gene expression control. This information can in turn be used to prioritize which factors should be the focus in rule-based optimization approaches or to modify feature sets that are fed into machine learning algorithms to discover more nuanced effects.

In this study, we generated three codon-randomized mRFP sets using an approach based on Golden Gate assembly, yielding a total of 1459 reliable, high-quality data points. We then used these data points (pairs of CDSs and expression values) as training data for our machine learning algorithm MEW (mRNA

Expression Wizard) to establish an algorithm that can predict the expression from the CDS. From inferred feature importances it is concluded that the bases surrounding the start codon and the RBS are far more important determinants of gene expression than codon usage throughout the gene.

4.2 Results and Discussion

4.2.1 A method for full-gene codon randomization

To randomize synonymous codon usage throughout the whole *rfp* CDS, we developed a new randomization method based on Golden Gate assembly to link partially degenerate double-stranded DNA blocks. We used this approach to generate three codon-randomized mRFP sets. The first set ($\mathrm{CAI_M}$) is fully randomized and uses an equal distribution of all synonymous codons for each amino acid. This approach results in a uniform codon bias distribution across the gene, with an expected average CAI of 0.67 (Figure 4.1). To diversify the CAI within codon random sequences, we generated two additional libraries by restricting the allowed relative adaptiveness (the usage ratio of a codon to that of the most abundant synonymous codon). We made these sets to detect the effects of randomization in sequences comprised of common codons and sequences comprised of rare codons. The set with a low CAI ($\mathrm{CAI_L}$) used only synonymous codons with a relative adaptiveness <0.60 or the lowest relative adaptiveness in the case the synonymous codons are used in a close to equal ratio, resulting in an average CAI of 0.41. The set with a high CAI ($\mathrm{CAI_H}$) used only synonymous codons with a relative adaptiveness >0.50, resulting in an average CAI of 0.83. The theoretical number of possible sequences for the libraries are displayed in Table S4.1.

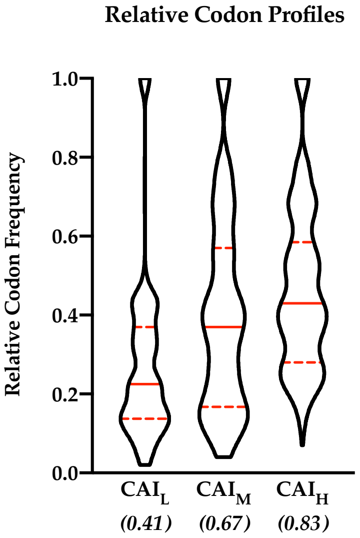


FIGURE 4.1: **Relative codon frequency profiles of the theoretical (IUPAC) $\mathrm{CAI_L}$, $\mathrm{CAI_M}$ and $\mathrm{CAI_H}$ libraries.** The solid red line indicates the median, dotted red lines indicate the quartiles. The CAI of each library is given in italics.

First, we converted the amino acid sequence of the mRFP gene into an IUPAC-formatted degenerate nucleotide sequence. Unfortunately, due to our synthesis method, it is impossible to cover the full set of synonymous codons for Arginine, Leucine and Serine using a single IUPAC notation. Therefore, the synonymous codons for these three amino acids were limited to 4 out of 6 possibilities to still cover the highest number of possibilities for each (CGN, CTN and TCN). Similarly, the three stop codons (TAA, TGA and TAG) can also not be annotated in a single IUPAC annotation, and therefore the stop codon was kept constant (TGA).

Next, we split the IUPAC encoded CDS into roughly equal parts of 85 nucleotides in such a way that each contained unique 4 base pair overlaps with neighbouring parts. In order to generate these fixed overhangs, some codons with multiple synonymous options needed to be fixed to a single codon. Together with the limitations of the UIPAC annotations, this resulted in a lower number of experimental sequence possibilities compared to the theoretical sequence space (Table S4.1). The DNA parts were ordered as single-stranded oligos with additional type 2S restriction sites flanking the blocks. The oligos were converted to double-stranded DNA using PCR and consequently assembled with a Golden Gate reaction (Figure 4.2a, b). Only a small fraction of the total DNA parts assembled into the full product of 707 bp (Figure 4.2b, indicated with the mRFP tag). Seven intermediate products were observed that did not further assemble into the full gene. The assembly limitations could originate from synthesis errors in the initial oligo, preventing type IIS restriction or resulting in incorrect overhangs. A single transformation of the libraries in *E. coli* (DH10B) yielded between 150.000 and 320.000 colonies, of which about 70% gave a detectable level of fluorescence (Figure 4.3). The remaining 30%, for which no or very little fluorescence was measured, mainly comprised constructs that had a frameshift in the ORF. This is not unexpected, as some blocks are likely missing one or multiple nucleotides since the coupling efficiency of oligos is not 100%. These errors eventually lead to frameshifts and thus protein truncations or mutations.

To enrich for low, medium and high expressing constructs, a preselection was performed using FACS for each library (Figure 4.2c). We picked an equal number of colonies from the FACS cell fractions from the three groups of each library, all of which were individually inoculated in liquid cultures. These cell cultures were measured, using flow cytometry for single-cell measurements and with a microplate reader for culture measurements. We amplified the mRFP-encoding DNA using colony PCR and amplicons were analyzed by Sanger sequencing.

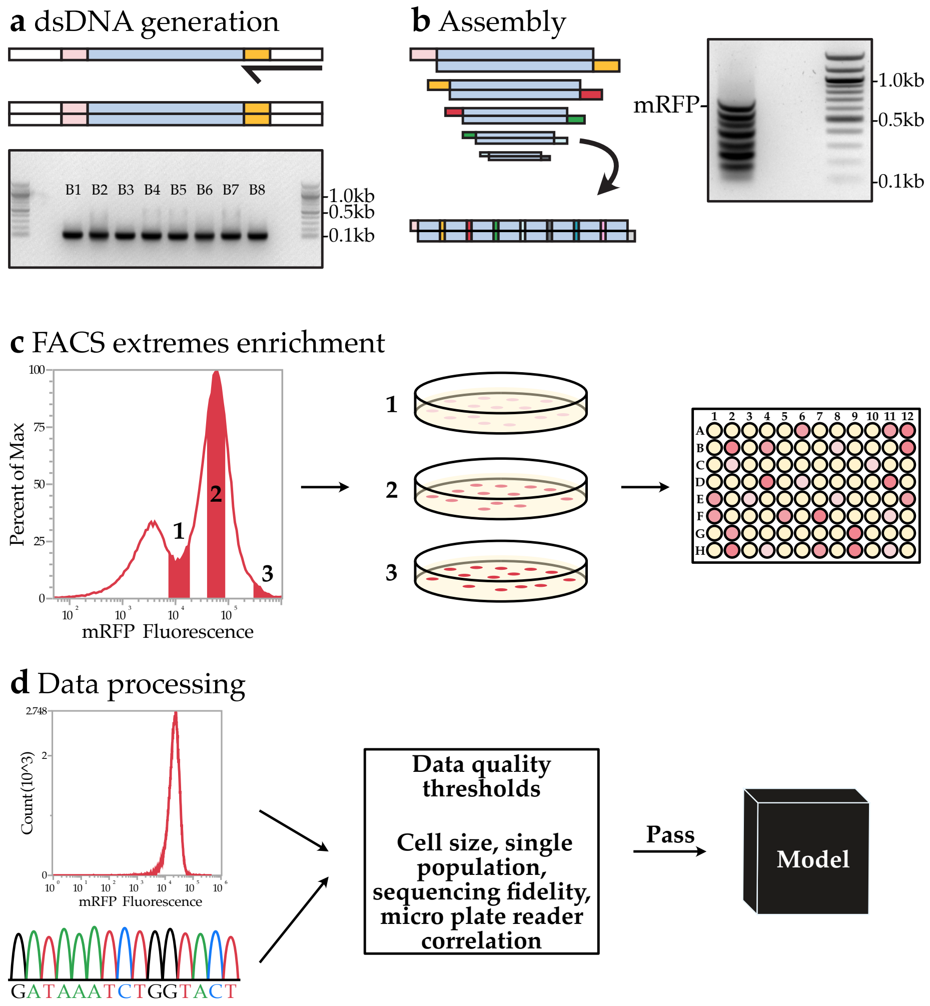


FIGURE 4.2: **Codon random library generation and analysis.** (**a**) Illustration of PCR to generate dsDNA from oligo and the electrophoresis gel showing eight blocks used to build codon random RFP. (**b**) Illustration of the assembly reaction and the electrophoresis gel result of the assembly. The complete assembly of all eight blocks is indicated. The seven bands below are intermediate products. (**c**) FACS enrichment for a wide expression range within the library to have a higher representation of the high and low expressing codon variants. (**d**) Flow Cytometry analysis of cultures and Sanger sequencing data are QA passed and used in computer learning models.

Next, the data was evaluated to exclude low-quality sequencing quality reads, mixed populations, and deviations in cell morphology (see Methods and Figure S4.1 as a few colonies formed enlarged or clumped cells). Additionally, the cell cultures were measured using a microplate reader to exclude a few rare data points where the expression deviates more than 25% from the average relationship between the two measuring methods (see Methods and Figure S4.2).

To identify the determinants of translation efficiency, and to assess if the expression levels could be predicted on the basis of the gene sequences, we employed two different machine learning approaches: Random Forest Regressor (RFR) and LASSO (Least Absolute Shrinkage and Selection Operator). For this purpose, we developed MEW: the mRNA Expression Wizard, which can train and test a variety of machine learning models using different types of featurisations. These featurisations include methods that focus on the base pair composition of the coding sequence to observe the effects of factors like translation elongation efficiency and vectorisations that do not look at base identity at all but only care about the probability that a base is paired in the context of an mRNA secondary structure.

Our rationale to use both LASSO and RFR is that due to the stepwise decision making, RFRs should be able to capture interdependencies between bases, while LASSO is better suited to straight-forward regression and feature selection. Importantly, for each regressor we trained, we extracted the feature importances to identify determinants of translation efficiency. We trained separate regressors for both full-length featurised mRNA sequences and for sliding windows of varying sizes along the entirety of the mRNA, to assess if certain windows are more predictive of translation efficiency than others.

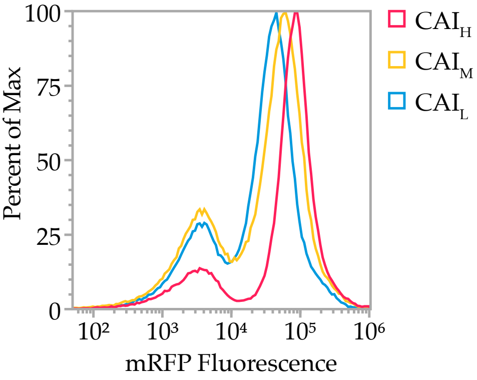


FIGURE 4.3: **Normalized flow cytometry overlay of the mRFP fluorescence signal from the CAI_L CAI_M CAI_H library.** The left peak is part of the population showing no fluorescence mainly due to assembly errors in the CDS. The right peak shows the mRFP expression of each library. The average expression of the CAI_L CAI_M CAI_H libraries is in increasing order but high expressing variants are found in all libraries (right tail). The ratio between the left and right peaks shows the fidelity of the library as the left peak consists of autofluorescence of non-expressing or non-functional variants.

ML Algorithm	Featurisation	Pearson correlation (p-val)		Spearman correlation (p-val)	
LASSO	BPP	0.485	(0.000)	0.497	(0.000)
	one-hot	0.754	(0.000)	0.753	(0.000)
	BPP + one-hot	0.773	(0.000)	0.775	(0.000)
RFR	BPP	0.519	(0.000)	0.525	(0.000)
	one-hot	0.777	(0.000)	0.772	(0.000)
	BPP + one-hot	0.753	(0.000)	0.750	(0.000)

TABLE 4.1: Pearson and Spearman correlations for LASSO and RFR machine learning models trained on the full mRNA sequence using different featurization methods.

As machine learning algorithms are only as good as the data that they are given, featurising our mRFP data in a way that captures most information was key. We used three featurization methods: one based on one-hot encoding which looks at base identity, another based on predicted pairing probabilities of bases in mRNA secondary structure calculated with ViennaRNA (Lorenz et al., 2011), and a third which combines these two. While one-hot encoding should in theory also be able to capture base pairing probability, we decided to also use mRNA secondary structure featurization as the interpretability of the resulting features is a lot greater. We will call the three types of featurisations BPP (base pairing probability), one-hot, and BPP + one-hot respectively.

4.2.2 Translation efficiency can be predicted from mRNA sequence

We trained and validated our RF and LASSO regressors using 10-fold cross-validation, yielding a value of predicted translation efficiency for each data point. When the algorithms were trained on the entire length of the mRNA sequence, this gave rise to significant Pearson correlations between actual and predicted expression data for all combinations of featurisations and algorithms (Table 4.1, Figure 4.4).

The performances of LASSO and RFR were comparable, independent of the featurization method used. Notably, BPP encoding underperformed consistently compared to one-hot encoding. One reason for this may be the uncertainty of the predictions of ViennaRNA: because the secondary structures and base pairing

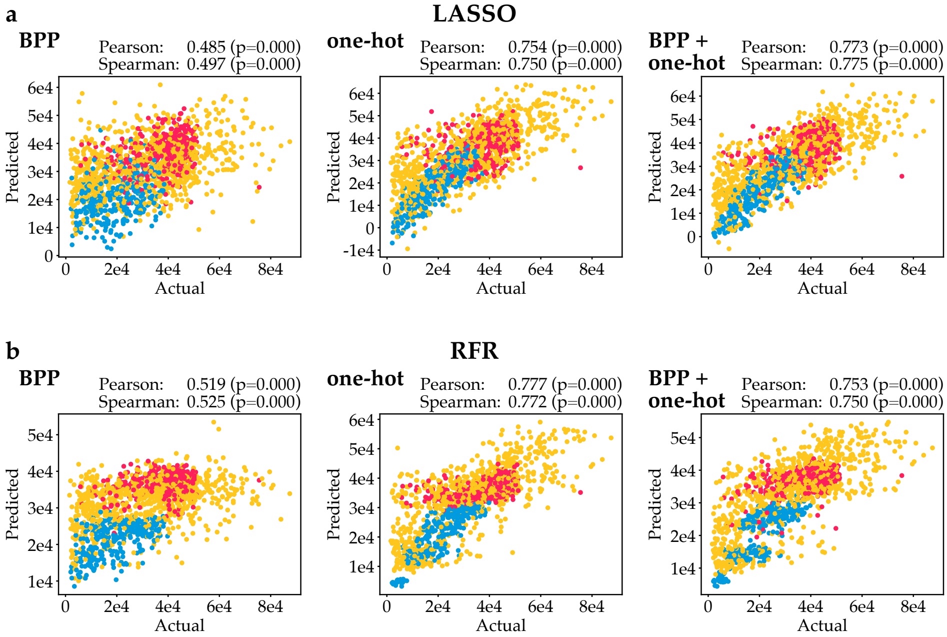


FIGURE 4.4: **Actual expression data vs predicted expression using various machine learning algorithms and featurisations of full-length mRNA.** Blue, yellow and red points indicate data points from the $\mathrm{CAI_L}$, $\mathrm{CAI_M}$ and $\mathrm{CAI_H}$ libraries respectively. (**a**) Actual expression data vs predicted expression using LASSO. (**b**) Actual expression data vs predicted expression using RFR.

probabilities that were computed for each molecule have a degree of uncertainty, the features that the algorithm is trained on may not be a completely accurate representation of reality. In contrast, one-hot encoding captures absolutely all information in the mRNA molecule and therefore is expected to perform better. Nevertheless, algorithms trained with BPP featurization still yield predictive regressors that predict translation efficiency much better than random, indicating that mRNA secondary structure alone is an important determinant of translation efficiency. Finally, BPP + one-hot encoding did not seem to outperform one-hot encoding alone. This is in line with our expectation that, in principle, all information on base pairing probability should already be captured by one-hot encoding. Therefore, while BPP encoding may be more interpretable than one-hot encoding for humans, this extra information is redundant for machines.

4.2.3 Bases surrounding the start codon and the RBS are most predictive of translation efficiency

Next, we assessed which features, and by extension which bases, are most predictive for translation efficiency. We did this by extracting the coefficients for LASSO, and the feature importances for RFR, and plotting them against sequence position (Figure 4.5). We found that, independent of the featurizations and algorithms used, the most predictive bases were always close to the start of the CDS, including the 5-10 bases before the start codon and the first 25 bases following the start codon. In comparison, the rest of the codons play a minimal role in predicting translation efficiency. This strongly suggests that mRNA secondary structure plays a dominant role in determining translation efficiency, and not tRNA availability on which measures such as CAI rely. Interestingly, this agrees very well with previous studies which found the same effect for GFP and attributed this to the necessity for an unobstructed RBS (Kudla et al., 2009; Cambray, Guimaraes, and Arkin, 2018). However, since we used a BCD system which should in theory reduce RBS obstructions prior to ribosome binding by straightening out the mRNA with a leading ribosome, the importance of the area surrounding the RBS is unexpected in our study. This might indicate that smaller, transient interactions within mRNA molecules are capable of quickly and reversibly forming secondary structures, even after the leading ribosome has passed, which can still lead to partial obstruction of the RBS and thus reduced translation efficiency.

Interestingly, the LASSO regressor using BPP for featurization gives positive coefficients to the bases immediately succeeding the RBS2 (Figure 4.5a, first panel). As a positive coefficient indicates that involvement in mRNA secondary structures at this position is positively correlated with gene expression levels, this may seem counterintuitive. Possibly, the RBS is more accessible for the ribosome if the region directly downstream is involved in a (weak) secondary structure, as this may prevent the RBS itself from base pairing with the bases directly downstream. Note that the bases in the RBS2 itself and the 5′ of the coding region are overwhelmingly assigned negative coefficients. In concordance, the presence of purines in the 5′ of the coding region, particularly 'A', is strongly correlated with protein production levels, while the presence of pyrimidines tends to be negatively correlated with protein production levels in this region (Figure 4.5b, panel 1). Possibly, because the RBS2 contains exclusively purines (A and G), the presence of purines in the 5′ coding region reduces the chance of secondary structure formations.

To further substantiate our finding that bases surrounding the start codon dictate translation efficiency, we trained regressors on sliding windows of 10, 20, 30,

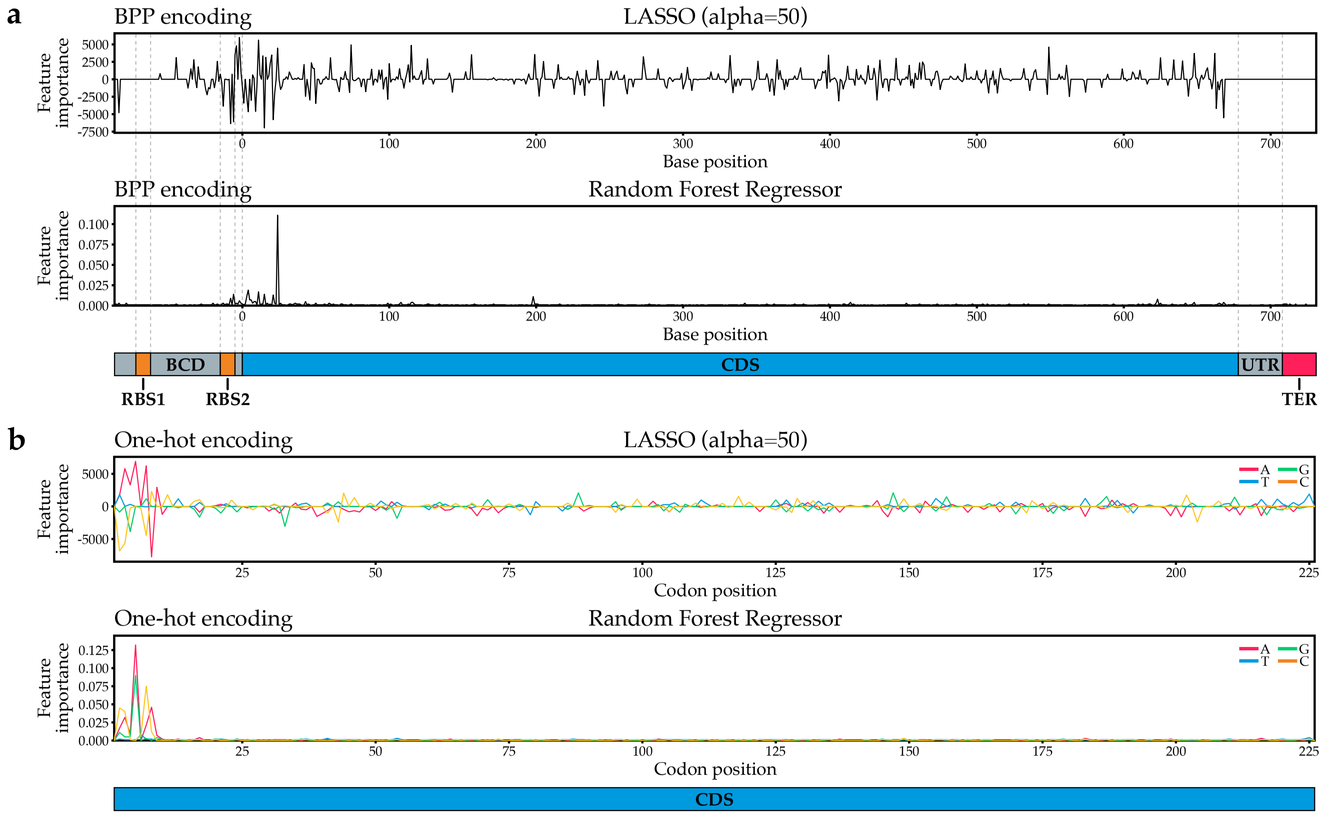


FIGURE 4.5: **Continues on next page**

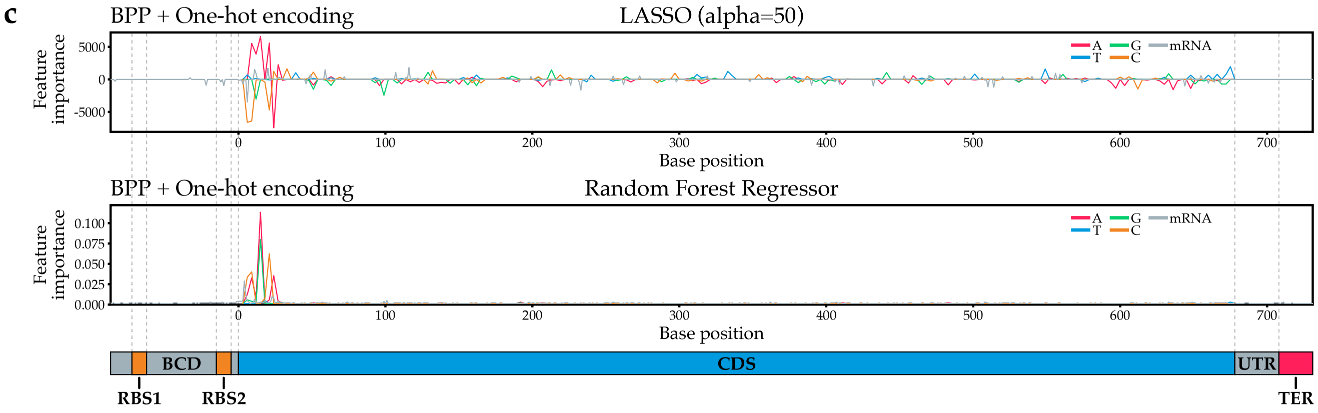


FIGURE 4.5 (CONT.): **Feature importances for various machine learning algorithms and featurisations.** LASSO feature importances are coefficients: a positive coefficient indicates a positive correlation between a base and translation efficiency, a negative coefficient indicates a negative correlation. In RFR, feature importances are always positive and therefore they say nothing about the directionality of the correlation. (**a**) Feature importances for algorithms using BPP featurisation. (**b**) Feature importances for algorithms using one-hot encoding. Since only every third one-hot encoded base of the coding sequence varies, only every third base of the coding sequence was plotted. (**c**) Feature importances for algorithms using BPP + one-hot featurization.

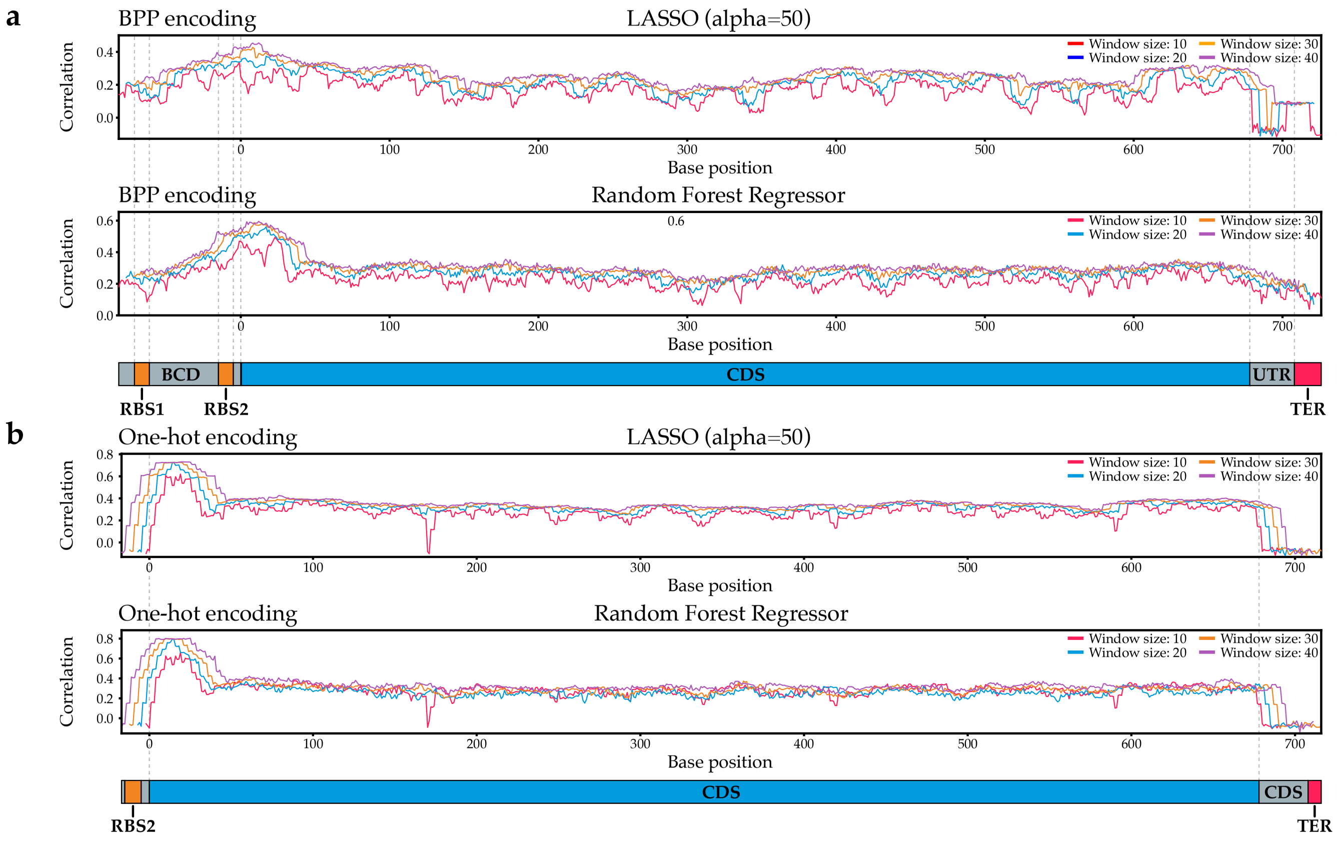


FIGURE 4.6: **Continues on next page**

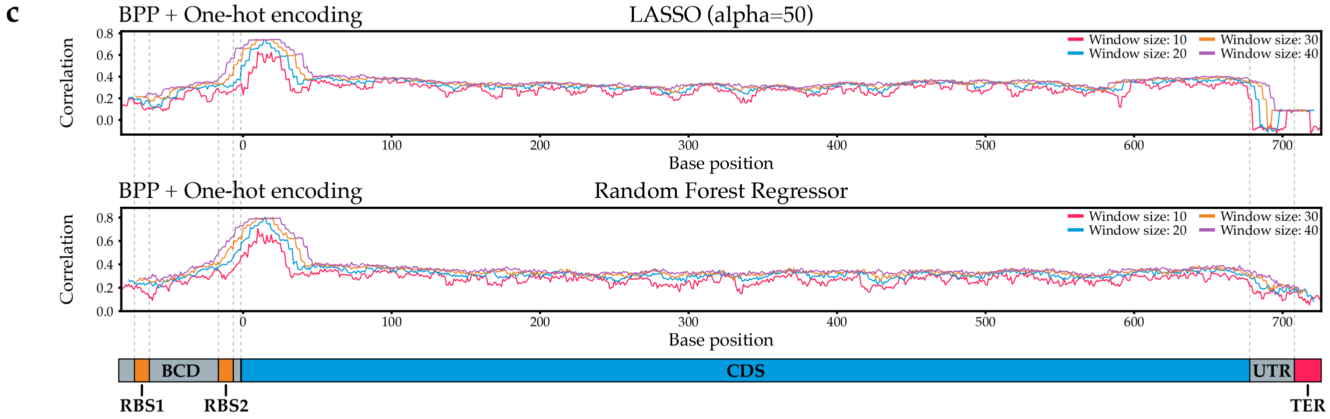


FIGURE 4.6 (CONT.): **Predictive regions of translation efficiency in mRFP mRNA.** The x-axis represents the central base of a sliding window of indicated lengths, and the y-axis the correlation between actual expression data and the expression that was predicted by a machine learning algorithm trained on solely that sliding window. (**a**) Predictive regions found by algorithms trained with BPP featurization. (**b**) Predictive regions found by algorithms trained with one-hot encoding. As the one-hot encoded features for the UTRs are constant and thus contain no information, windows only containing residues in the UTR were omitted. (**c**) Predictive regions found by algorithms trained with BPP + one-hot featurization.

or 40 base pairs to visualize which regions of the mRNA were most predictive of translation efficiency. This approach ensures that we do not miss any effects further on in the mRNA molecule that may correlate to effects observed in the 5′ region. For each sliding window, we performed 10-fold cross-validation and plotted the correlation between actual expression data and the predicted expression data against the position of the sliding window (Figure 4.6). Clear peaks of increased predictive power can be observed around the start codon, which corroborates our earlier findings that it is this region that is primarily responsible for dictating translation efficiency. This is especially apparent in models trained with one-hot encoded features and BPP + one-hot encoded features. Specifically, the 20 nucleotides surrounding base 15 are necessary for optimal prediction accuracy (Figure 4.6b, c, window size 20). In fact, the RFRs trained for this window perform better than the RFRs trained on the entire sequence: the Pearson correlation between actual and predicted expression data for this window lies about 0.05-0.1 higher than the Pearson correlation for the entire sequence. This increase in performance is not observed for the corresponding window in LASSO, even though this window does still perform better than other windows for all featurization methods used. However, this is to be expected because random forest regressors are more sensitive to overfitting than LASSO. Hence, by only selecting the relevant window, a lot of dataset-specific noise is filtered out. The observation that most of the predictive power originates from the 5′ end of the CDS aligns with the previous observation by Kudla *et al*. They used GFP as a reporter with a T7 promoter and concluded that region -4 to +37 could explain 44% of the variation in fluorescent levels. (Kudla et al., 2009).

We also observed some ‘dips’ in performance. One such dip can be seen for small window sizes in the 3’UTR when base pairing probabilities were used as features in LASSO regressors (Figure 4.6a, c). This region is very invariable both in terms of sequence and secondary structure: since the terminator almost always forms an incredibly strong secondary structure, the bases directly before it will almost never be involved in a secondary structure. As a result, the features representing this region hold no information at all. The effect is exacerbated for small windows, as they are less likely to capture predictive residues upstream or downstream of an information-devoid region. In contrast, the secondary structure of the terminator itself does appear to be slightly informative. An explanation could be that sometimes the terminator might form a secondary structure with a different region in the mRNA, causing improper transcription termination due to interference in the stem-loop formation or reduced mRNA stability. However, it is important to keep in mind that correlations between actual and predicted expression data for regressors trained on this region are still extremely low. Therefore, while the 3'UTR region holds some information, it is not likely to be very influential.

A second dip is located around base 165 and 166 for regressors using one-hot-encoded featurisations (Figure 4.6b). This information valley is caused by an unusually constant region in the mRFP gene, particularly in the $\mathrm{CAI_H}$ set, due to the low codon variability of local amino acids and a fixed boundary region of two assembly blocks. This is an artefact of our method, and hence not a biologically relevant observation. This dip is not observed for featurization methods that also include base pairing probabilities, and this makes sense: while the identity of bases in that region may be invariable, the mRNA secondary structure can still vary due to interactions with upstream or downstream regions.

To better understand which sequence elements in the 20-window surrounding base 15 affect translation efficiency, we plotted feature importances for each regressor trained on this window (Figure 4.7). From this, we inferred that especially at position 15, low probabilities of involvement in mRNA secondary structure are predictive of high expression. This is in line with the current consensus that minimal mRNA secondary structure surrounding the 5′ end of the coding region is conducive to efficient translation. In the case of mRFP, this low base pairing probability seems to be primarily achieved by placing an 'A' at position 15 (Figure 4.7b, c, codon 5).

For our best-performing regressor, we plotted actual expression data against predicted expression for each data point through cross-validation. This revealed a very strong correlation ($r = 0.803$) for all three libraries (Figure 4.8), and demonstrates that mRFP protein production can be predicted extremely well by just looking at bases 6-25 of the entire coding sequence. In the case of GFP, similar conclusions have been made. (Kudla et al., 2009).

Finally, we compared three of the best expressing variants from our random library to several modern or commonly used codon optimization methods: The AG29G and FCFOall method, developed by Boēl *et al.* based on their so-called model M (Boël et al., 2016); the codon optimization method offered via the popular platform Benchling, based on DNA Chisel (Zulkower and Rosser, 2020); the commonly used commercial algorithm from the DNA synthesis company GeneArt (Fath et al., 2011; Raab et al., 2010); and two older but still frequently cited algorithms JCAT (Grote et al., 2005) and OPTIMIZER (Puigbo et al., 2007). All three variants, which originated from the $\mathrm{CAI_M}$ set, displayed better expression than all these previously reported methods (Figure 4.9a), this while our library

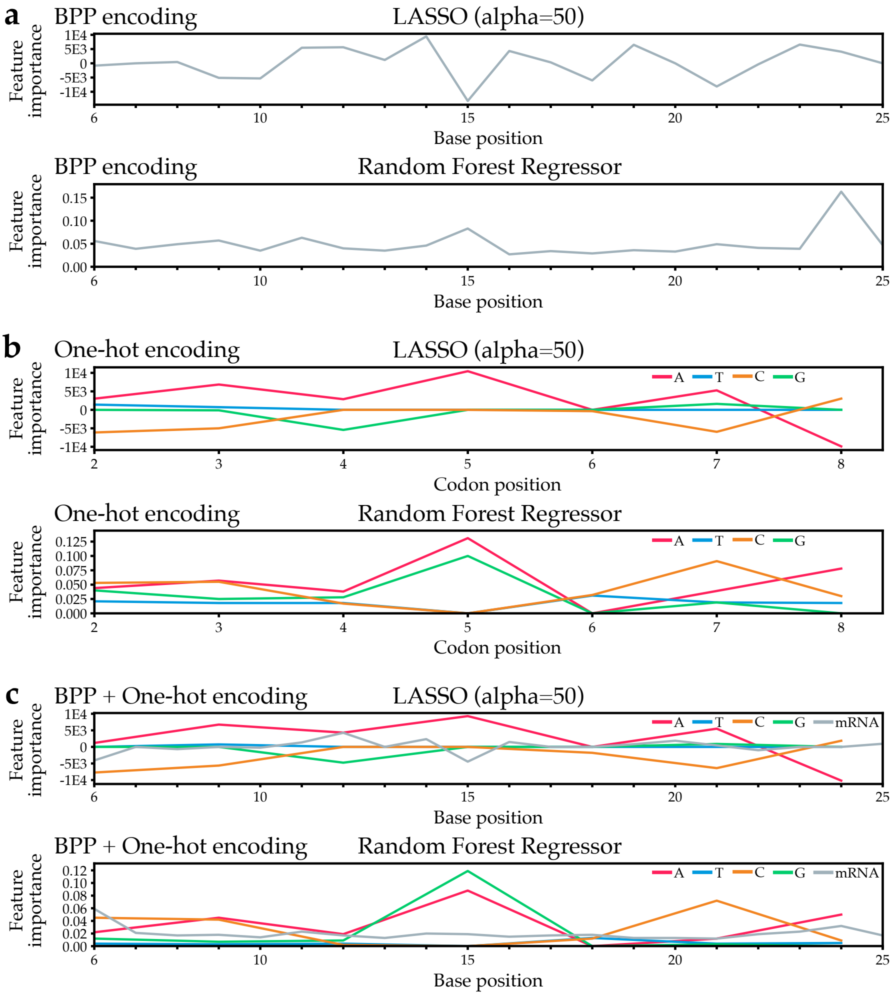


FIGURE 4.7: **Feature importances for various machine learning algorithms and featurisations trained on a 20-bp window around base 15.** LASSO feature importances are coefficients: a positive coefficient indicates a positive correlation between a base and translation efficiency, a negative coefficient indicates a negative correlation. In RFR, feature importances are always positive and therefore it contains no information about the directionality of the correlation. (**a**) Feature importances for algorithms using BPP featurisation. (**b**) Feature importances for algorithms using one-hot encoding. Since only every third one-hot encoded base of the coding sequence varies, only every third base of the coding sequence was plotted. (**c**) Feature importances for algorithms using BPP + one-hot featurization.

represents only a fraction of the total possible sequence space and the three sequences contain at least 30% rare codons (with a relative codon frequency < 0.5). This indicates clearly that there is still room for improvement.

Regressor type	RFR
Featurisation type	BPP + one-hot
Window size	20
Window location	Base 15
Pearson COR	0.803
Spearman COR	0.798

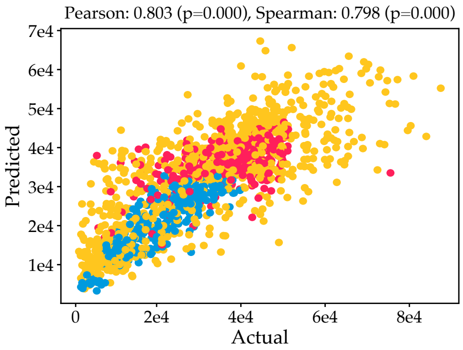


FIGURE 4.8: **Actual expression data vs predicted expression for our best-performing regressor.**

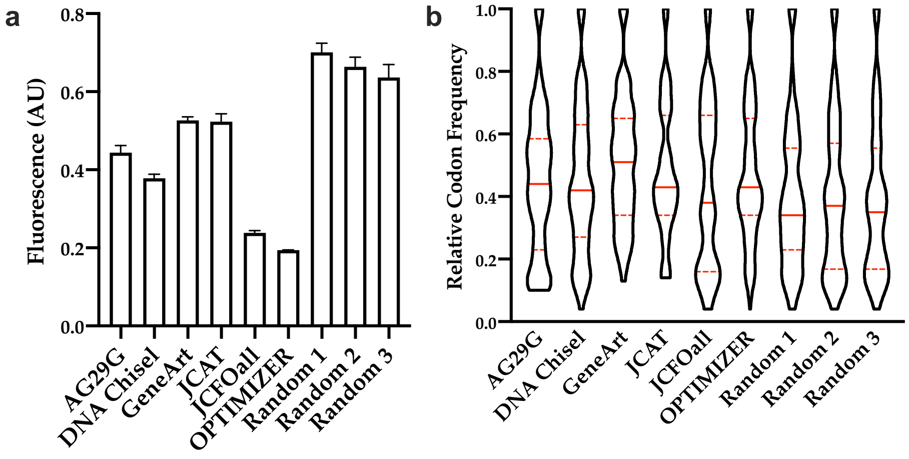


FIGURE 4.9: **Benchmark results.** (**a**) Fluorescence measurements of different codon optimized mRFP constructs compared to the top 3 expressing constructs found via codon randomization. (**b**) Relative codon frequency profiles of the benchmark genes and the top 3 expressing construct found via codon randomization. The solid red line indicates the median, dotted red lines indicate the quartiles.

4.3 Conclusion

We have generated a well-performing predictive algorithm using a random forest regressor algorithm. However, the predictability of this algorithm is likely limited to the studied protein: mRFP. Since the majority of the predictability originates from our One Hot featurization, the algorithm cannot account for the effects of the fixed first and second nucleotide of the codons, which is amino-acid-dependent, and thus protein-dependent. The algorithm's performance was less when only the BPP featurization was used. However, this approach is more universal and can be applied to other proteins. Overall, the algorithm highlighted the bases 6-25 as the most information-dense region. This means that the majority of the expression can be explained by the codon usage within this region. We expect that the codon identity is linked to secondary structure formations with the 5′UTR and that this is generally a stronger determinant of protein production than specific codon usage, although we cannot rule out the influence of e.g. the previously mentioned ramp or other unknown factors in the 5′CDS. Despite the predictive information hot-spot in the 5′CDS, overall codon usage does contain some predictive information (Figure 4.5) and on average we saw that a high CAI leads to higher expression than a low CAI (Figure 4.3). A likely explanation for this is that, because the translation initiation is the major rate-limiting factor, the effects of codon usage are less apparent for most sequences. Only when codon usage is optimal in the 5′CDS and the translation initiation rate is high, general codon usage becomes more relevant. As also argued by Kudla *et al.*, the causality between optimal codon usage and high protein production might be reversed (Kudla et al., 2009). Only when translation initiation rates are high codon usage might become relevant to reduce tRNA depletion. This is also highlighted by the fact that we found our highest performing variants not in the CAI_H library but in the CAI_M library. This library contains a lot more variation (Table S4.1) and therefore the chance of generating an accessible 5′UTR and optimal 5′CDS increases.

Despite the use of a bicistronic design our expression was still mostly related to codon sequence identity at the 5′CDS. Hence, we were not able to extract the more subtle rules regarding overall codon usage. A good next step would be to exclude the influential 5′CDS region from randomization. This then allows the effects of overall codon usage to better translate into overall protein output.

An interesting approach towards codon optimization for applications requiring high protein production would be to randomize the influential 5′CDS region and optimize the remaining codons using an advanced algorithm. Ideally, this algorithm is developed using an optimal fixed 5′CDS so it optimizes only for translation elongation. By using *in vivo* selection methods (Rennig et al., 2018) the optimal 5′CDS can then be selected and, due to the codon optimization method, the burden of this highly expressed gene on other cellular processes is limited.

It is clear that still not all features influencing translation efficiency are completely understood but it seems that the majority of protein production variation due to codon usage originates from codon usage in the 5′CDS. Assuming that this is a generic feature for all genes expressed in prokaryotic production systems, this would put current approaches that are used for optimizing gene expression, which often evolve around using optimal codons to reflect tRNA availability, in a different perspective. Based on our analysis, we should not be focusing on codon optimization along the entire length of the mRNA, but instead, focus our attention primarily on redesigning the region surrounding the RBS. Many of the current codon optimization algorithms do not require the 5′UTR sequence for their prediction. Based on here presented results, we strongly recommend that, for *in silico* optimization of protein production, these codon optimization algorithms (especially for protein production in bacteria) are adjusted to take into account the importance of the 5′UTR and the 5′ end of the CDS. Alternatively, as described here, an experimental approach is used in which a smaller randomization region (5′UTR + 5′codons 1-10) is generated and screened for high expression.

4.4 Materials and methods

4.4.1 mRFP codon randomization

The amino acid sequence of the monomeric Red Fluorescent Protein (mRFP) was used to generate three IUPAC-annotated DNA sequences representing our libraries (CAIL, CAIM and CAIH). Each IUPAC annotated sequence was then split into blocks of roughly equal sizes (80-90 nucleotides) in such a way that they have a unique 4-base pair overlap with their neighbouring blocks. Overhangs were selected from a set that is optimized for high ligation fidelity (Potapov et al., 2018). To create each required overhang, we attempted to fix degenerate codons in such a way that the separate blocks were roughly equal in size, and that loss of degeneracy stayed limited. For example, fixing the degenerate sequence ARAT to AAAT would result in the loss of 1 codon possibility, while fixing the degenerate sequence YGCN to CGCC would result in the loss of 7 codon possibilities. 5′ and 3′ flanking sequences containing recognition sites for the Type II restriction enzyme BsaI-HF®v2 (NEB, R3733) were added to each IUPAC DNA block, to generate the unique single-stranded overhang after digestion. The 5′ of the first block and the 3′ of the last block contained SapI (NEB, R0569) recognition sites instead. Each block was ordered as a DNA oligo (Ultramer® DNA Oligonucleotides, IDT) and using a strand-displacing Taq polymerase (NEB, M0482), the ssDNA was converted to double-stranded DNA via PCR. PCR reactions containing the dsDNA block were cleaned and concentrated to 20 µL mQ using the DNA Clean & Concentrator™-5 kit (Zymo, D4004). 4 µL Gel Loading Dye Purple (6x) (NEB, B7024) was added to each block and they were loaded on a 1% agarose gel and ran for 30 minutes at 100 volts. The dsDNA blocks are excised from the gel and purified to 20 µL mQ using the Zymoclean™ Gel DNA Recovery Kit (Zymo, D4002). 5 µL of the dsDNA was used to quantify the DNA concentration with the Qubit assay (Invitrogen, Q32853) according to the manufacture's protocol.

The dsDNA blocks were mixed in an equal molar ratio to a total volume of 41 µL, with 5 µL T4 Ligase Buffer (NEB, B0202) 400 units T4 Ligase (NEB, M0202) and 60 units BsaI-HF®v2 (NEB, R3733). Assembly reaction was done overnight at 37 °C for 18 hours, followed by 5 minutes at 60 °C and a holding step of 12 °C. The assembly is cleaned and concentrated to 15 µL mQ using the DNA Clean & Concentrator™-5 kit (Zymo, D4004). 3 µL Gel Loading Dye Purple (6x) (NEB, B7024) was added and the assembly mixture was loaded on a 1% agarose gel and ran for 40 minutes at 100 volts. The full-length assembled product was excised from the gel and purified to 44 µL mQ using the Zymoclean™ Gel DNA Recovery Kit by Zymo (D4002). 10 units of SapI (NEB, R0569) was added with 5 µL CutSmart Buffer (NEB, B7204) and digested for 2 hours at 37 °C. The digested codon random mRFP with single-stranded overhangs was cleaned and concentrated to 15 µL mQ using the DNA Clean & Concentrator™-5 kit by Zymo (D4004). The complete 15 µL containing the codon random mRFP library is used in a ligation reaction to generate the plasmid library.

4.4.2 Plasmid preparation and library generation

The pFAB3909 plasmid (Mutalik et al., 2013a) (Addgene #47812) with a P15A origin, kanamycin resistance gene and bicistronic design element was modified to be able to accept the codon random mRFP library and include a constitutive GFPuv expression. The relatively weak bla promoter was used to drive the mRFP expression, keeping the total protein yield relatively low for high producing mRFP codon variants preventing expression saturation. A strong terminator was used for efficient transcription termination and to enhance mRNA stability. The open reading frame was replaced by SapI recognition sites to generate the sticky overhangs that accept the mRFP library and a large part of nonsense DNA was inserted between the SapI sites to separate the double SapI digested plasmid from linear product. A GFPuv gene, driven by the P4 promoter, was added to the plasmid as an internal standard for gene expression. Expression of GFPuv is weak as to not interfere with the mRFP expression efficiency but strong enough for detection with flow cytometry.

About 3 µg plasmid was digested with 20 units SapI (NEB, R0569) and dephosphorylated with 3 units rSAP (NEB, M0371) with 6 µL CutSmart Buffer (NEB, B7204) in a total volume of 60 µL for 3 hours at 37 °C, followed by an inactivation step at 65 °C for 20 minutes. The linear plasmid was excised from the gel and purified to 30 µL mQ using the Zymoclean™ Gel DNA Recovery Kit by Zymo (D4002). The codon random mRFP library (15 µL) was ligated into 30 ng linear plasmid with 400 units of T4 ligase (NEB, M0202) and 2 µL T4 Ligase Buffer (NEB, B0202) in a total volume of 30 µL for 18 hours at 16 °C. The ligation mixture was cleaned and concentrated to 10 µL mQ using the DNA Clean & Concentrator™-5 kit by Zymo (D4004). 1 µL of the codon random mRFP library is transformed into electrocompetent DH10B cells (20 µL competent cells, 2mm cuvette, Voltage: 2500V, Resistor: 200 Ω, Capacitor 25 µF, BTX® ECM630). Cells were recovered in 1 mL NEB® 10-beta/Stable Outgrowth Medium (NEB, B9035) at 37 °C for 1 hour. The cells were transferred to a 50 mL tube and 9 mL LB (10 g/L Peptone (OXOID, LP0037), 10 g/L NaCl (ACROS, 207790010), 5 g/L Yeast Extract (BD, 211929)) was added with 50 µg/L kanamycin (ACROS, 450810500) and incubated for 18 hours at 37 °C.

4.4.3 Expression range enrichment and selection

A FACS (Sony, SH800S Cell Sorter; GFPuv excitation at 488 nm, emission at 525/50 nm; mRFP excitation at 561 nm, emission at 617/30 nm) was used to sort 50.000 cells of the overnight cell culture into 3 groups based on expression. The left and right tail of the normal distribution and a part of the middle peak was sorted to create 3 groups of low, medium and high expression. The 3 cell groups were put on individual agar plates (10 g/L Peptone (OXOID, LP0037), 10 g/L NaCl (ACROS, 207790010), 5 g/L Yeast Extract (BD, 211929), 15 g/L Agar (OXOID, LP0011), 50 µg/L kanamycin (ACROS, 450810500)) and grown overnight at 37 °C. From these plates individual colonies were picked and grown in 2 mL 96 well plates with 200 µL LB with kanamycin (10 g/L Peptone (OXOID, LP0037), 10 g/L NaCl (ACROS, 207790010), 5 g/L Yeast Extract (BD, 211929), 50 µg/L kanamycin (ACROS, 450810500)) for 18 hours at 37 °C.

4.4.4 Measurements and sequencing

The cell cultures were diluted 100x in PBS (8 g/L NaCl (ACROS, 207790010), 200 mg/L KCl (ACROS, 196770010), 144 mg/L Na2HPO4 (ACROS, 12499010), 240 mg/L KH2PO4 (ACROS, 447670010)). mRFP expression was measured using a flow cytometer (Thermo, Attune NxT Flow Cytometer; GFPuv excitation at 405 nm, emission at 512/25 nm; mRFP excitation at 561 nm, emission at 620/15 nm; stop option 200.000 single cells). A gate was used to exclude GFPuv outliers (±10% of the total population) and thus consequently unrelated biological variance was reduced as the GFPuv expression level are expected to stay constant. From the overnight cultures 1 µL of cells were used in a PCR reaction to obtain the gene for Sanger sequencing using Q5 (NEB, M0492). The PCR reaction was sent to Macrogen Europe B.V. for sample clean-up and Sanger sequencing.

4.4.5 Data validation

For each final datapoint, a few criteria have to be met. The expression data needed to show a clear single population. We excluded cultures that showed a double peak graph for the mRFP fluorescence or showed a noticeable difference in cell size. The raw sequence data was validated by extracting the open reading frame sequence using in-house scripts. If all bases in the open reading frame had a Phred quality score > 20 (a base call accuracy of at least 99%) and the translated sequence matched the mRFP amino acid sequence the sequence passed and was used in the analysis. Finally, all cell cultures were also measured using a microplate reader (BioTek, Synergy Mx). 50 µL overnight cell cultures were diluted in 50 µL PBS (8 g/L NaCl (ACROS, 207790010), 200 mg/L KCl (ACROS, 196770010), 144 mg/L Na2HPO4 (ACROS, 12499010), 240 mg/L

KH2PO4 (ACROS, 447670010)). The plates were incubated at room temperature for 1 hour before measuring (cell density measured at 600 nm; GFPuv excitation at 395/9 nm, emission at 508/9 nm; mRFP excitation at 584/9, emission at 607/9 nm). The microplate reader fluorescent readings were normalised with the OD600 for both the GFPuv and mRFP readings. If the relation between the microplate reading and flow cytometry reading varied more than 25% from the overall relation the datapoint was discarded. For the remaining data points, we assessed dataset-wide biases and correlations, such as assembly bias and the correlation between expression level and GC content to ensure the dataset as a whole was appropriate for machine learning.

4.4.6 Featurisation for machine learning

In order to prepare mRNA sequences so that they could be used as input features for machine learning, they were vectorised in two ways. The first method one-hot encodes each third base of an mRNA sequence as a vector of length four, beginning at the start codon and ending at the stop codon. We only one-hot encoded each third base as our randomization strategy ensured that the first and second bases of each codon remained constant. For mRFP coding regions, which consist of 226 codons each, this yielded feature vectors of length 904 (226 x 4). The second approach uses the RNAFold utility from the ViennaRNA package (python interface, v2.4.14). With RNAFold, a base pairing probability matrix which stores the pairing probabilities for each pair of bases was calculated for the complete mRNA sequence, including the upstream bicistronic design (BCD) and the downstream terminator. For each base, the pairing probabilities in its row were summed to yield a vector where each entry represents the probability that a base is paired. This gave rise to vectors of length 818, corresponding to one probability for each base of the BCD, coding sequence, and terminator. We also fed combined feature vectors into our machine learning regressors by concatenating the vectors representing base pairing probabilities to the one-hot encoded vectors. The resulting vectors had a length of 818 + 904 = 1722. Finally, we divided the entire mRFP mRNA sequence, including the 5'UTR and 3'UTR, into sliding windows of 40 base pairs, and made feature vectors for each window based on base pairing probability, one-hot encoding, and the two combined, as described above. For base pairing probabilities within these windows, the entire mRNA molecule and not just the window was folded with ViennaRNA to capture long-range interactions as well.

4.4.7 Building machine learning regressors

To assess if mRNA expression level could be predicted from sequence, we employed two different machine learning approaches: Random Forest Regressor (RFR) and LASSO. We implemented RF and LASSO using the scikit-learn package (v0.23.0, ref) in python (v3.7.6), with the sklearn.ensemble.RandomForestRegressor and sklearn.linear_model.Lasso modules respectively. For RF, default settings were used, while for LASSO, nine different values for alpha were assessed (1.0, 2.0, 5.0, 10.0, 20.0, 50.0, 100.0, 200.0 and 500.0; max iterations = 10,000). An alpha of 50.0 performed best for one-hot encoding and base pairing probability encoding separately, while an alpha of 100.0 gave rise to the best predictions for featurization that combines both encodings. Separate regressors were constructed for full-length featurised mRNA sequences and for each sliding window of 10, 20, 30, or 40 base pairs. Regressor accuracies were evaluated through 10-fold cross-validation where 90% of the data are used to predict the translation efficiency of the other 10%. This is done for each 10% of the data, such that we have a predicted translation efficiency, measured with flow cytometry, for each data point. From these predictions, Pearson and Spearman correlations were computed for actual flow vs predicted flow of the data points in the out groups. Feature importances were extracted from all ten regressors built in cross-validation, averaged, and plotted and visualized with matplotlib (v3.2.1). Code and regressors are made available at `https://git.wageningenur.nl/terlo012/mew`.

4.5 Supplementary information

Library	Theoretical sequence space	Experimental sequence space
CAI_L	4.47e49	6.22e38
CAI_M	3.19e104	3.68e93
CAI_H	2.01e53	4.50e48

TABLE S4.1: Theoretical and experimental sequence spaces for CAI_L, CAI_M and CAI_H libraries.

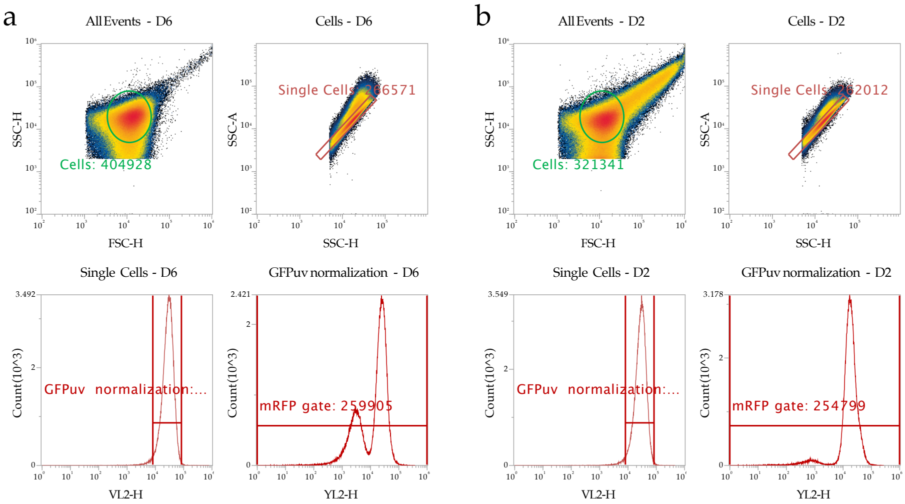


FIGURE S4.1: **Failed QA examples for double population and increased cell size/clumping.** (**a**) A flow cytometry environment where a clear double population is present in the mRFP expression (bottom right panel). (**b**) A flow cytometry environment where there is an increase in cell size or clumping of cells (top left panel, difference becomes apparent when compared to **a**). The reason for this phenomenon is unknown.

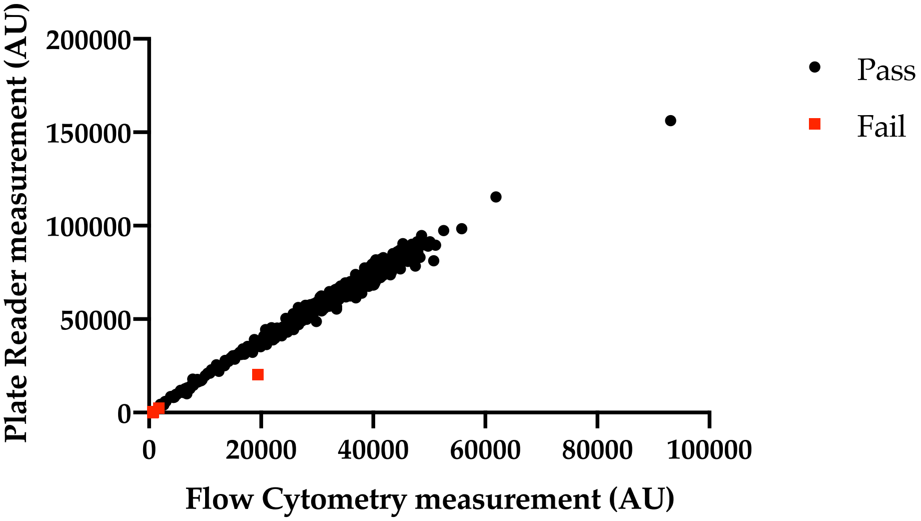


FIGURE S4.2: **Flow Cytometry data plotted against Plate Reader data.** Illustrative for the points that were excluded in the quality assurance check (in red). These points deviated more than 25% from the average ratio between all points.

Chapter 5

Translational feed-forward and feed-back control

Sjoerd C.A. Creutzburg*, **Thijs Nieuwkoop***, Thijmen Zegers & John van der Oost.

Laboratory of Microbiology, Wageningen University,
Stippeneng 4, 6708 WE Wageningen, The Netherlands

Abstract

Genes that are co-expressed as polycistronic mRNAs are not necessarily translated with the same efficiency. Differences in protein synthesis in such cases generally are caused by distinct rates of translation initiation and/or elongation, which in turn are governed by their ribosome binding site, their codon usage, and/or their mRNA secondary structure. Translational coupling of downstream genes and their upstream counterparts is a well-established feed-forward phenomenon, in which the translation of an upstream cistron influences that of a downstream one. In contrast, we here describe different types of feed-back control of gene expression. First, we demonstrate that a downstream gene may influence the expression of an upstream gene. In addition, we show a major impact of the sequence of the 3'UTR, including the spacer between the coding sequence and the terminator. Moreover, we show that the ratio between the translation of the genes in an operon is also dependent on the transcription rate. It is concluded that, even after half a century of intense research, the sequences of the translated and untranslated regions of genes and operons still have unpredictable impact on the relative rates of the transcription and translation processes, and hence are crucial determinants for the efficiency of gene expression. The here-presented results may contribute to elucidating the molecular basis of these phenomena, which is crucial for fundamental understanding as well as for applications that rely on operon design.

5.1 Introduction

Prokaryotes often generate polycistronic mRNA for the concerted transcription of functionally related genes, for instance for enzymes that compose metabolic pathways and for subunits of protein complexes (Galperin and Koonin, 2000; Huynen, 2000). While transcription of the genes clustered as an operon is generally equal, differences in translation rate may cause differential expression of these genes. Differential translation may arise from differences in ribosome binding efficiency, codon usage and impediments like strong secondary structures. Between successive genes in the operon, translational coupling has already been observed several decades ago (Oppenheim and Yanofsky, 1980; Schümperli et al., 1982; Aksoy, Squires, and Squires, 1984). More recently, a systematic and quantitative characterisation of *E. coli* operons has shown that the expression of an upstream gene can influence the expression of a downstream gene, depending on the length of the intergenic region (Levin-Karp et al., 2013). Increased translation of an upstream gene by incorporating a range of ribosomal binding sites (RBS), each having different ribosome binding strength, has a direct or indirect effect on the translation rate of the downstream genes. Furthermore, it was found that translational coupling is also affected by so-called polar mutations in the upstream gene's coding sequence (Oppenheim and Yanofsky, 1980). For instance, a point mutation in the upstream gene *trpE* affects the expression of the downstream *trpD*. Two main models have been proposed to explain this phenomenon.

The first model is based on ribosomal "flow-through". Ribosomes terminating translation at the upstream gene's stop codon are in the direct vicinity of the downstream gene's initiation sites, thus a direct feed-forward control may occur (Govantes, 1998). This model was further confirmed by increasing the distance between the stop codon and the downstream start codon (Levin-Karp et al., 2013). A decrease in translational coupling was found by increasing the length of the intergenic region. The second model is based on the helicase activity of the 70S ribosome. Increased translation of an upstream gene can dissolve secondary structures throughout the operon, potentially removing inhibitory structures present on the downstream gene's initiation region. Strong secondary structures upstream of the *atpA* gene's translation initiation region were indeed found to inhibit the translation rate of *atpA* (Rex et al., 1994). A model was proposed in which the secondary structure within the upstream *atpH* cistron is dissolved by the processive ribosome activity, also resulting in unfolding the downstream mRNA to improve accessibility for ribosomal binding to allow translation initiation of *atpA*. This translational feed-forward coupling phenomenon has important implications for designing operon synthesis as well as for operon reduction. The alterations of gene order and sequence not only affect translation rates of the altered gene but can also affect expression ratios throughout the operon, resulting in differential stoichiometries (Quax et al., 2013).

In this study, we explored the relation between local sequence mutations, secondary RNA structures and, consequently, gene expression levels throughout an operon and found a new form of coupling. We conclude that gene expression is not only influenced by upstream sequences as previously described (feed-forward control), but also vice versa by downstream sequences through feed-back control. Whilst keeping RBS sequences constant, differences in expression of the upstream gene can be observed when altering downstream gene sequences. The translation rate of an upstream monomeric red fluorescent protein (mRFP, hereafter called RFP) gene is altered when the downstream green fluorescent protein (GFPuv, hereafter called GFP) gene sequence is altered. The effect has been assayed of co-expressing an upstream gene encoding a single RFP variant with a downstream gene encoding one of four types of GFP (wild-type GFP, harmonised GFP, frameshifted GFP and a functional mutant of GFP). To have a homogenous transcript size, a strong rho-independent terminator was added, which increased protein production drastically. Moreover, substantial differences in expression were detected when the 30 bp linker between the stop codon and the terminator hairpin was varied (*post* stop, *ante* terminator region or PSAT region). The most obvious explanation for the observed fluctuations in gene expression is a change in stability of the corresponding mRNA, although the increase in gene expression does not match the increase in mRNA levels. Furthermore, the contributing effects of this region are not associated with a specific open reading frame sequence and therefore offer a new generic means of controlling gene expression.

5.2 Results

5.2.1 Translational coupling occurs regardless of ORF order in the mRNA

The difference in codon usage between the wild-type (WT) and harmonised (H) GFP-encoding gene (Claassens et al., 2017) causes a difference in expression level, where the harmonised gene is expressed significantly more than the wild-type. To allow accurate quantification of this effect, we cloned a monomeric red fluorescent protein (RFP) downstream the GFP variants as an internal control (Figure 5.1a; pTN001). The GFP fluorescence indeed shows that the harmonised GFP is better than the wild-type, but, unfortunately, the RFP fluorescence also fluctuates (1.09-fold; p = 0.0487), most likely reflecting translational coupling (feed-forward control). This implies that the rfp gene in these constructs cannot be used as an internal control. Reversing the order of the genes, *rfp* upstream and *gfp* downstream (Figure 5.1a; pTN002), RFP foremost lowers the expression of GFP in favour of RFP. However, instead of diminishing the translational coupling, expression of the *gfp* gene influences that of the *rfp* gene even more when it is located downstream of the RFP. To further investigate this reciprocal translational coupling (feed-back control), a GFP frameshift mutant (FS) with a 4-nucleotide insertion halfway, and a GFP functional mutant (FM) with a Tyr66Ser mutation that prevents the formation of the fluorophore, were made (Figure 5.1b; pTN002). In addition to this set, in order to make all of the transcripts the same size, a strong synthetic terminator was inserted downstream of the GFP with a 45-nucleotide spacing sequence between the stop codon and first nucleotide of the terminator stem. This spacing sequence was generated by a random number generator and selected for approximately 50% GC content and lack of secondary structure as predicted by mFold (Zuker, 2003). While the WT-GFP showed significantly higher GFP fluorescence, we failed to obtain correct clones with the terminator behind the H-GFP; we only obtained clones with mutations in the *h-gfp* coding region or in the terminator stem and of a *h-gfp* gene disrupted by the insertion of a transposon. Given that the H-GFP has more expression than the WT-GFP and the terminator increases expression, the bacteria most likely could not cope with the burden or internal GFP concentration in the case of the H-GFP construct. Therefore, this set of constructs was abandoned.

A new set of constructs was made with, instead of the P_{tacI} promoter, the weaker P_{bla} promoter (Deuschle et al., 1986) controlling the operon without a terminator (pTN003) and with terminator (pTN004). In the absence of the terminator, the weak promoter diminishes the expression of both GFP and RFP to very low levels (Figure 5.1b; pTN003), while including the terminator (Figure 5.1b; pTN004) restores the H-GFP expression almost to the level of the PtacI promoter (Figure 5.1b; pTN002). When the pTN003 and pTN004 constructs are normalised to their respective WT-GFP constructs (Figure 5.1c, d), it becomes clear that the interdependency of GFP, RFP and their surroundings is severe. Figure 5.1c shows the weak promoter without a terminator. The large error bars in the GFP is because the expression of GFP is so low that the total fluorescence is only 10% above that of the frameshift variant (FS-GFP) and just 4x the fluorescence background of the medium. The RFP has no detectable auto-fluorescence from either the medium or the cells, so it can still be measured accurately at low expression levels. Since the expression of GFP (and RFP) is higher with the terminator, fluorescence of the pTN004 construct (Figure 5.1d) can be measured accurately.

Regardless of the constructs' promoter or presence of a terminator, the basic pattern for the translational coupling is the same. The RFP expression is highest

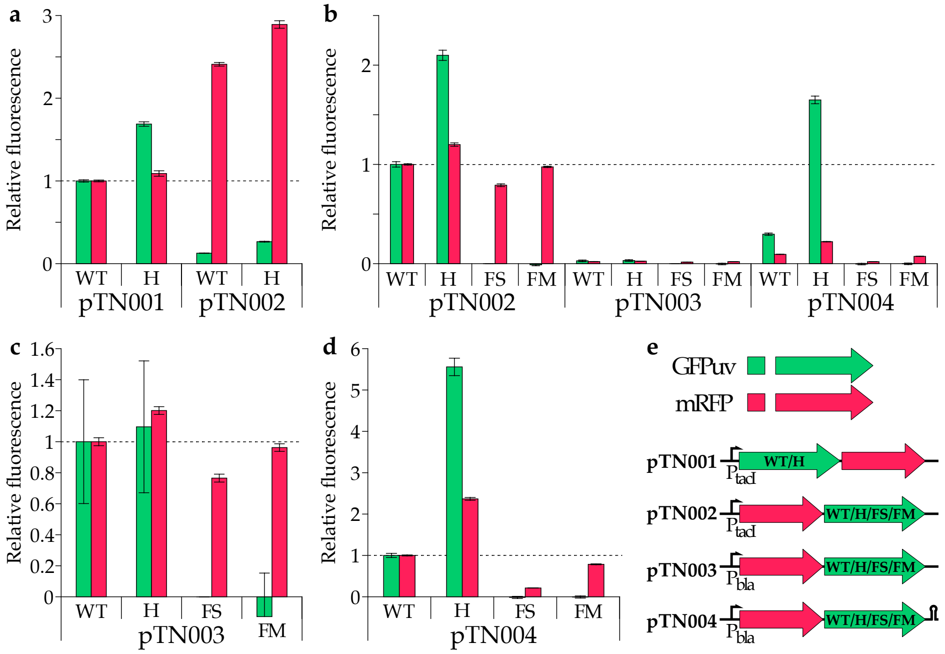


FIGURE 5.1: **Coupled expression of RFP (red) and GFP (green) in the same operon.** **(a)** A small increase in RFP can be seen with a better expression of GFP. pTN001 has the GFP in front of the RFP, while pTN002 has it reversed. The expression of both GFP and RFP is highly dependent on their respective position in the operon, but in all cases, the wild-type GFP (WT) shows less fluorescence than the harmonised GPFuv (H). Error bars represent the standard deviation. **(b)** Compared to pTN002, the pTN003 and pTN004 have a weak promoter. The pTN004 has, in addition, a terminator behind the GFP. The weak promoter diminishes the expression of pTN003 severely and can be partially compensated for by the addition of the terminator in pTN004. The terminator enhances the stability of the mRNA, but not all genes profit from that to the same extend. The RFP/GFP ratio is vastly reduced in pTN004 compared to pTN002, while the H/WT ratio is increased. **(c)** pTN003 dataset of (b) normalised for pTN003-WT. **(d)** pTN004 dataset of (b) normalised for pTN004-WT. **(e)** Construct overview of the pTN001-4 plasmids. P_{tacI} is about 17 times as strong as P_{bla}. Cloning a terminator behind pTN002 was unsuccessful.

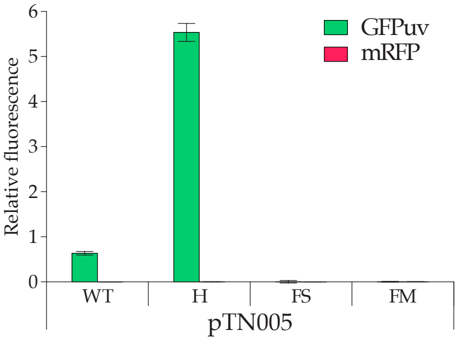


FIGURE 5.2: **Fluorescence data for pTN005.** The GFP variants are under control of the P_{bla} promoter and transcription is terminated by a terminator. Hence, fluorescence values are normalised for pTN004. Green fluorescence of the WT is less than in pTN004, while the others are roughly the same. Red fluorescence is totally absent, excluding the GFP contributing to red fluorescence in the other constructs.

when co-expressed with H-GFP, followed by WT-GFP, FM-GFP and lastly FS-GFP. This strongly suggests the occurrence of translational coupling independent of the gene order, either feed-forward (Figure 5.1a) or feed-back (Figure 5.1c, d). The WT-GFP and FM-GFP have almost the same sequence, as is reflected in the expression level of RFP. The FS-GFP also has almost the same sequence as the WT-GFP, but the key difference is that the frameshift mutant yields a truncated GFP (half the size of the WT) due to a premature stop codon. The terminator appears to amplify the effect of translation efficiency of GFP and the coupled RFP. The pattern of GFP in pTN004 is more pronounced compared to pTN002 (Figure 5.1b); pTN002 H-GFP/WT-GFP is just over 2, while pTN004 H-GFP/WT-GFP is over 5 (Figure 5.1b, d). GFP fluorescence of pTN003 (Figure 5.1c) is only 10% of the background fluorescence, so the H-GFP and WT-GFP are not significantly different. Comparing the RFP expression, while the actual RFP expression is vastly different, compared to their respective WT-GFP, the pTN002 constructs and pTN003 constructs are very similar (pTN002 - Figure 5.1b; pTN003 – Figure 5.1c; RFP). The constructs of pTN004 (Figure 5.1d) show amplification of that pattern. The amplification is possibly also true for the FM-GFP, but since its RFP expression is similar to that of, or perhaps slightly lower than, the WT-GFP, it is difficult to ascertain.

To exclude the possibility that the fluorescence at 607 nm (normally attributed to RFP) is influenced by GFP directly, we made the constructs with a Pbla promoter, GFP only and a terminator (pTN005). The GFP fluorescence is comparable to the GFP fluorescence in pTN004, while the fluorescence in the RFP spectrum cannot be detected at all (Figure 5.2), showing that GFP itself cannot be responsible for changes in the fluorescence at 607 nm.

5.2.2 The effect of the 3′UTR on expression

We then looked into the major effect on expression by the addition of a synthetic terminator. RT-qPCR analysis of the pTN003 and pTN004 constructs (Figure 5.3a) shows a clear stabilising effect of the terminator. The terminator increases the mRNA abundance during the mid-log phase by approximately a factor of 2, regardless of codon use. Contrary, the increase in protein production (measured as fluorescence) is significantly higher than that (Figure 5.3b), with a clear influence on the codon usage. The H-GFP fluorescence increases by a factor 50 when adding a terminator, while the WT-GFP increases "only" a factor 10. The RFP increases by a factor of 4 and 9 respectively. Since the codon use has no significant effect on the mRNA concentration (Figure 5.3a), it is not likely that ribosome shielding through more efficient translation contributes to enhanced mRNA stability.

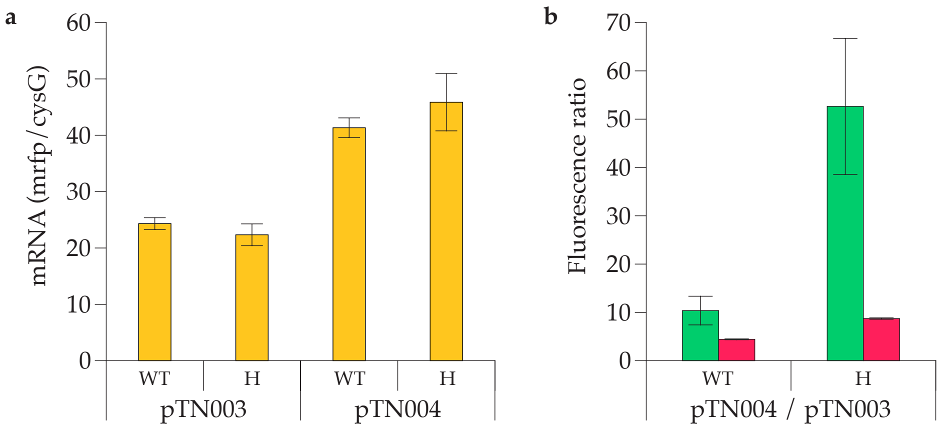


FIGURE 5.3: **RT-qPCR data for pTN003 and pTN004 compared to the fluorescence. (a)** mRNA abundance was estimated by qPCR on the RFP gene. The internal standard cysG (Zhou et al., 2011) was used for normalisation. **(b)** Fluorescence ratio of pTN004 and pTN003.

Next, we looked into a possible effect on expression by the 45-nucleotide PSAT region in the 3′UTR. To this end, a GFP and an RFP library were generated containing a completely randomised 30 bp PSAT region, replacing the 45 bp original. The length of PSAT is based on a recent study that reported that a sequence smaller than 30 nucleotides can have a negative effect on the terminator's termination efficiency, while increasing the size above 30 nucleotides did not show any

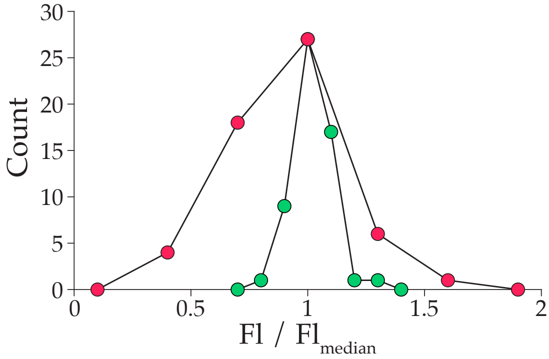


FIGURE 5.4: **PSAT region library.** Green indicates the PSAT sequences found for GFP, and red are the PSAT sequences for RFP.

effect (Li et al., 2016). After randomisation and transformation of the plasmid libraries to *E. coli*, transformant cells with a range of fluorescence were obtained for both RFP (5.4-fold difference) and GFP (2.7-fold difference) (Figure 5.4). The associated PSAT region sequences were obtained via Sanger sequencing (Table S5.2). A selected sequence set was interchanged between the two reporters to determine whether the effect of the PSAT sequence on the fluorescent level is protein-specific (pTN006 series; Table S5.1). Five sequences with a representing fluorescent range were selected for both the RFP and the GFP library (Figure 5.5a, PSAT 1-10). The sequences were cloned behind the original CDS, to serve as a control, and behind the alternative CDS. In addition, both sets of five were cloned behind *lacZ*, for an independent verification (Figure 5.5a). In an attempt to elucidate which part of the sequence is responsible for the high translation efficiency of PSAT region 9, a series of truncations were made from both the 5′ and 3′ end. Truncation even to 15 bp from either side does not appear to change the translation efficiency by much. Only the PSAT region 12, which ends in CCC, lowers the expression of all reporter proteins, indicating that interactions with the terminator might play a role. Surprisingly, there is a very good correlation (Figure 5.5b-d) between the observed relative GFP, RFP and LacZ levels, suggesting that the effect of the PSAT sequence is not CDS dependent, but rather a generic phenomenon.

5.2.3 Operon intergenic regions

The effect of the 3′UTR on translation may be interesting for tuning the expression of the genes in operons. In addition to tuning through sequences at the 3′ end of the operon, we set out to analyse the effect of the same PSAT sequences between the two coding sequences in the operon. Hence, a set of constructs was made with an intergenic region (IGR) either derived from PSAT region 1 (low translation; IGR-1) or PSAT region 10 (high translation; IGR-10) (Figure 5.5a;

pTN007 series; Table S5.1). Since it is unclear whether the terminator is involved in the aforementioned modulating effects, a version with and without a terminator stem was designed. IGR-0 is the control without any addition to the intergenic region, and the 3'UTR is the same as in construct IGR-16. RFP, WT-GFP and H-GFP were used in different orders (Figure 5.6). The fluorescence was normalised for the average of both of the IGR-0 constructs (either a GFP variant or RFP in the first position). While GFP and RFP values can be compared, the values do not represent an equivalent in protein molecules.

Remarkably, when no PSAT is inserted into the IGR (IGR-0), the first position is no longer favoured (Figure 5.6a, b; Figure 5.6d, e). This is in sharp contrast to the operons as depicted in Figure 5.1a, where expression of the gene at the first position is highly favoured. An explanation for this discrepancy might be differences in transcription rates, where high transcription (*tac* promoter, pTN001/pTN002) might cause limited ribosome availability. In that case, the first gene can already be translated while the second gene is not even transcribed, causing relatively high expression of the first gene in these operon constructs (Figure 5.1a), and much less so in case of the less efficient *bla* promoter (pTN004 and derivatives, Figure 5.1b-d, Figure 5.6). In the case of the WT-GFP, the order does influence the total expression of GFP and RFP. For example, this may be caused by secondary structures around the ribosome binding site associated with *gfp*. Both IGR-1 and IGR-10 in WT-GFP-[IGR]-RFP (Figure 5.6a) show a discrepancy between the poorly translated WT-GFP and the more efficiently translated RFP, indicating loss of translational coupling. Including neither IGR-1 nor IGR-10 suggest a strong influence on the translational coupling, but the substantial impact it has on the overall translation indicates that the coupling does persist. On the other hand, the constructs in the reverse order (RFP – WT-GFP; Figure 5.6b) do not exhibit this behaviour. The introduction of IGR-1 has the predicted effect on RFP translation, but WT-GFP translation appears to be largely unaffected. Extending the IGR with the terminator stem lowers the translational coupling somewhat, but not nearly as much as is seen in the WT-GFP-RFP constructs. Interestingly, the WT-GFP translation in several of the operons exceeds the translation of only WT-GFP (compare to Figure 5.6c).

Using the H-GFP instead of WT-GFP (Figure 5.6d, e) results in highly increased GFP fluorescence. The pattern of H-GFP – RFP is similar to the WT-GFP – RFP, but far less pronounced. In contrast, the RFP – H-GFP constructs have a more pronounced pattern. WT-GFP translation acts as a rate-limiting factor, so a poor IGR cannot attenuate the translation much further (Figure 5.6b). This is not the case for the H-GFP (Figure 5.6e), so the poor IGR1 becomes the limiting factor.

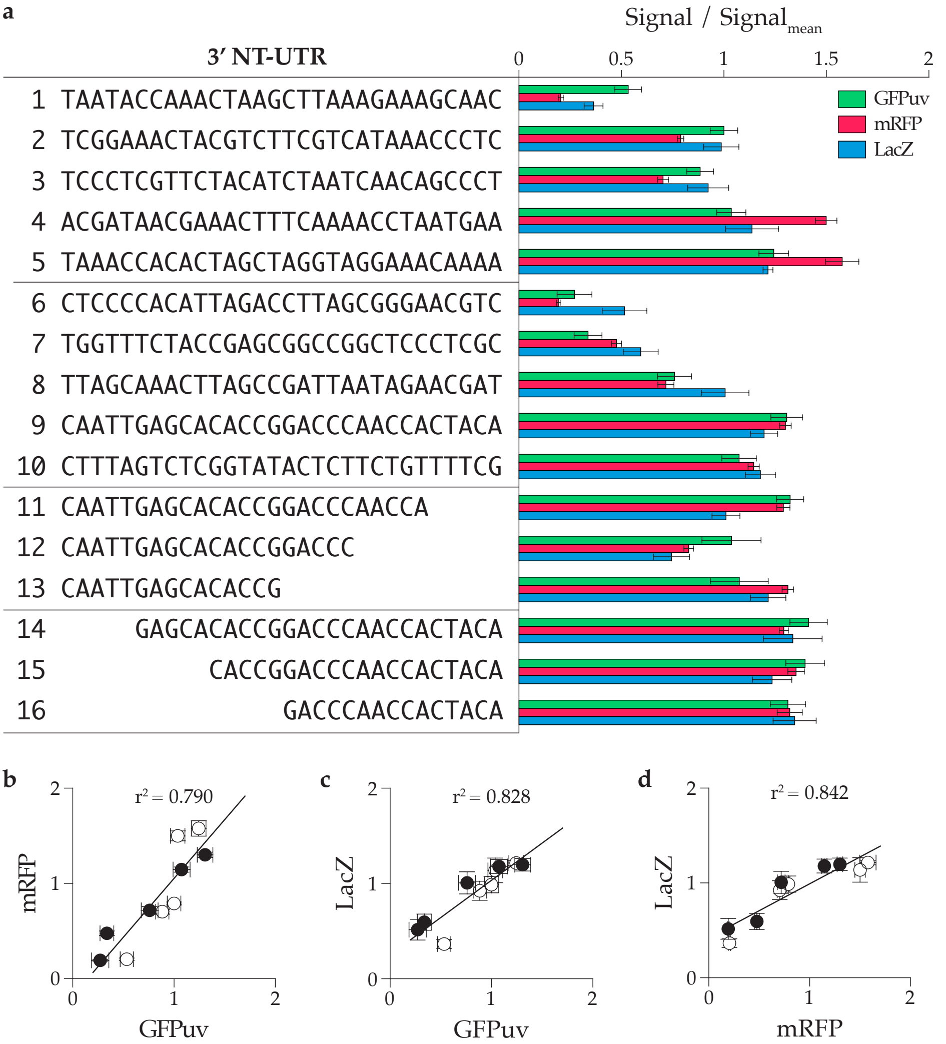


FIGURE 5.5: **Selected PSAT sequences and truncations of PSAT-9 with GFP, RFP and LacZ. (a)** PSAT sequence and relative signal. For GFP and RFP fluorescence was measured, for LacZ the hydrolysis of ONPG and the extinction at 420 nm. **(b-d)** Correlations between GFP/RFP, GFP/LacZ and RFP/LacZ. All have decent correlation, indicating that the effect the 3′UTR has on translation is mostly ORF independent.

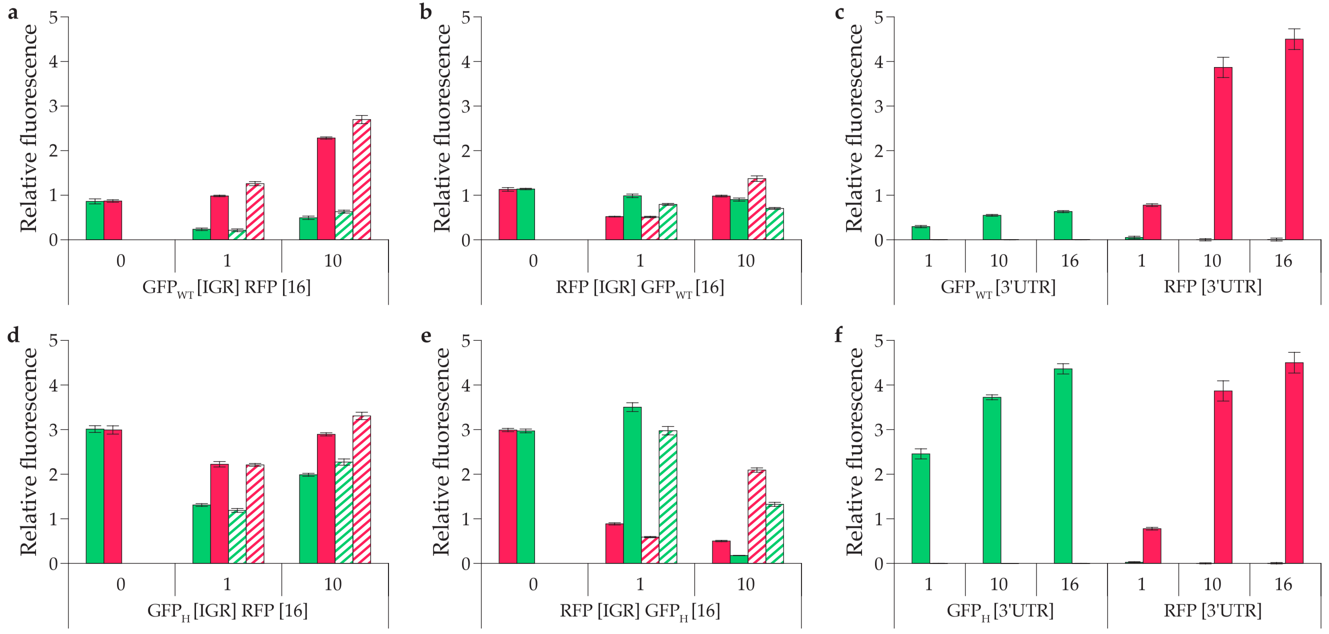


FIGURE 5.6: **GFP and RFP fluorescence in an operon with different intergenic regions (IGRs).** The numbers on the x axis (0, 1, 10, 16) indicate the PSAT region added to the IGR directly behind the stop codon (sequences as in Figure 5.5), where 0 is a control without any extra nucleotides. Solid: No terminator stem behind the PSAT region. Dashed: With terminator stem behind the PSAT region. Green: GFP fluorescence. Red: RFP fluorescence. All fluorescence has been normalised for the raw average of WT-GFP IGR-0. Error bars indicate the SD. **(a)** The WT-GFP followed by an IGR and RFP. **(b)** RFP followed by an IGR and WT-GFP. **(c)** Controls with either WT-GFP or RFP. **(d)** The H-GFP followed by an IGR and RFP. **(e)** RFP followed by an IGR and H-GFP. **(F)** Controls with either H-GFP or RFP.

5.3 Discussion

We here describe that the expression of genes in an operon can be coupled regardless of the order of those genes, either through feed-forward or feed-back control by both coding and non-coding sequences. The addition of a strong terminator at the 3′-end had a major influence on the expression of both GFP and RFP. Moreover, varying a spacer sequence (PSAT region) between the stop codon and terminator resulted in up to a 7.7-fold difference in expression, which appears to be largely ORF independent. Translational coupling has previously been observed in operons where the translation rate of the first gene influences the expression of at least two downstream genes (feed-forward) (Levin-Karp et al., 2013). It was explicitly stated that the accumulation of the downstream encoded protein had no influence on the upstream translation. In the literature, two models for translational coupling have been proposed. The first model is based on the disruption of secondary structures (Rex et al., 1994; Mossey and Das, 2013). Translation initiation can be severely hampered by a secondary structure masking the RBS. The helicase activity of the ribosome can disrupt these structures, but only when it is already bound (Takyar, Hickerson, and Noller, 2005). If the coding region of gene A forms an RNA structure with the RBS of gene B, thereby preventing ribosome binding, the frequency of ribosomes passing ORF-1 influences the availability of the RBS for the translation initiation of ORF-2. In feed-forward translation coupling (first ORF-1, then ORF-2), the secondary structure consists of two parts that are in relatively close proximity. An extreme example of this principle is the bi-cistronic design, in which a strong first RBS attracts a ribosome, after which the ribosome translates a short peptide that overlaps with the second RBS. The second RBS is now available for ribosomes to bind (Mutalik et al., 2013a). The second model for feed-forward translational coupling argues that the ribosome is in close proximity to the second RBS when it is released from the mRNA after the first gene has been translated. This model is supported by the introduction of a premature stop codon in the first gene diminishing the translational coupling, where the distance between the stop codon and the SD of the second gene is negatively correlated with the translational coupling (Schümperli et al., 1982; Levin-Karp et al., 2013). A combination of both models appears to be in good agreement with feed-forward translation coupling observations.

Based on the here presented experimental data, it is tempting to propose a model as well for the feed-back translational coupling phenomena. The aforementioned ’proximity model’ does not work for feed-back translational coupling. The downstream gene stop codon could be in the proximity of the upstream gene RBS if their respective 3′ UTR and 5′ UTR were to interact with each other, basically forming a loop between the beginning and end of the transcript. This is virtually impossible for the constructs in this study, because the part of the 5′ UTR available for this interaction is short (28 nt) with a very low GC content (18%). We observed that the introduction of a premature stop codon – via a frameshift – also decreased the expression of the upstream gene. Barring the long-range interactions between the beginning and the end of the transcript, this observation seems to contradict this model.

A more likely hypothesis is that the H-GFP construct is less prone to degradation by the degradosome due to increased ribosome shielding (Vytvytska et al., 2000; Braun, 1998; Edri and Tuller, 2014; Deneke, Lipowsky, and Valleriani, 2013). What determines the translation yield, is a combination of several factors during initiation (the RBS availability) and elongation (many factors, including the transcript's codon usage and the availability of matching charged tRNAs, and the transcript's secondary structure (Quax et al., 2015)). The combination of these factors may cause a huge difference in coverage of the mRNA by the ribosomes, up to a 100 fold (Oh et al., 2011; Ingolia et al., 2009). In contrast, the effect of mRNA stability appears to be limited and the codon usage has no significant effect on it (Figure 5.3).

The addition of a strong terminator improves the mRNA stability, as it forms a stem-loop which improves mRNA stability at the 3′ end protecting it from 3′-5′ exonuclease attacks (Newbury et al., 1987), and has been shown to increase gene expression 2-fold in *E. coli* (Vasquez et al., 1989) and over 30-fold in HeLa cell lines (West and Proudfoot, 2016)). We have found a rough approximation of a 4- to 50-fold increase in gene expression with the addition of a terminator, but that improvement is much larger than the improvement in mRNA concentration (lower than 2-fold). This indicates that the terminator does not only improve the stability of the mRNA, but also improves the translation rate of the mRNA. This is corroborated by the fact that the H-GFP profits much more from the terminator than the WT-GFP does. If the lack of a terminator causes a bottleneck after ORF-2, ribosomes stack up into ORF-2, unable to progress, and then the effect of codon usage in ORF-2 is diminished. This is likely also the case in pTN001 and pTN002, where the number of ribosomes may be limiting the translation, due to the high transcription of the operon. The terminator releases the bottleneck and the ORFs that were most hampered by this bottleneck profit the most. It is hard to see how the terminator located after ORF-2 directly influences the translation of ORF-1, since the effect on mRNA stability is so limited. A possibility is that the translation of ORF-1 is directly influenced by the translation of ORF-2. The best explanation would be a continuous ribosome train. Normally, after peptide chain termination, the ribosome stays bound to the mRNA awaiting recycling (Kiel, Kaji, and Kaji, 2007). The ribosome is about 20 nm, while an RNA base spans about 0.34 nm. This means that the ribosome may cover as many as 60 nucleotides. The intergenic region is rather short (42 nt), so likely, the main source for ribosomes for ORF-2 is the recycled ribosomes of ORF-1 (proximity effect). On the other hand, a bottleneck in ORF-2 is directly transduced to ORF-1. The distances are so short that a ribosome in the process of being recycled can block the way of ribosomes in ORF-1. Ribosome profiling (Ingolia et al., 2009) may reveal whether there are bottlenecks after the stop codon of ORF-2 that are simply not transcribed because of the terminator.

Randomising the PSAT region resulted in variable fluorescence levels. A sequence that gives rise to high GFP expression also results in relatively high RFP expression and vice versa. Secondary structures between the ORF and the PSAT region are unlikely to be the reason for the variability in fluorescence. It is more likely that the PSAT region interacts with the terminator in some way, either strengthening or weakening the stem structure, or that it affects how the ribosome behaves during the last stages of translation. We analysed 160 sequences for nucleotide preference for each position after the stop codon, and searched for patterns comparing the top tier and the bottom tier. However, no correlation could be found. Hence, the exact mechanism is still unknown at this point, but it appears that this region influences gene expression rather independently (the same sequence has a similar effect on three unrelated preceding ORFs, and therefore is highly interesting in protein production studies that require high protein yield. Since we do not know by what mechanism the sequences of the 3′ UTR (and by extension the IGR) influence the translation rate, it is difficult to explain their rather irregular behaviour in polycistronic mRNA. From what we gathered, a rapidly translated ORF can be severely hindered by a poorly translated ORF both upstream and downstream. Poor codon usage of the upstream ORF and a poor choice of IGR sequence cause translational impediment of the downstream ORF, which can be alleviated by increasing the distance (Figure 5.6a, d). Poor codon usage of the downstream ORF impedes the upstream ORF translation severely regardless of the IGR and the impediment is not solved by increasing the distance moderately (30-45 bp) (Figure 5.6b). However, the choice of IGR still influences the upstream ORF. The addition of the terminator stem to the IGR has a very moderate effect as well.

Altogether the results here show the considerations that must be taken into account when designing and studying polycistronic mRNAs. Besides the previously established forward translational coupling, we now show that reverse, feed-back translational coupling exists. A knockout of a single gene within an operon will affect the expression of both upstream and downstream genes, which might result in a phenotype that cannot be exclusively attributed to the absence of said gene. Instead of knocking out the gene, we advise opting for a functional mutant. If the gene's active site is unknown frameshifts could be introduced, however early stop codons should then be avoided.

5.4 Materials and methods

5.4.1 Strains and media

Throughout this study we used *E. coli* DH10B T1^R (Invitrogen C6400-03). Bacterial cultures used for cloning were grown in LB medium (10 g/L Bacto peptone, 5 g/L yeast extract, 10 g/L NaCl in demineralised water), with 50 mg/L kanamycin when appropriate. An additional 15 g/L agar is added for standard medium plates. The fluorescence assays were performed on M9TG (1x M9 salts (Sigma), 10 g/L tryptone, 5 g/L glycerol), which has allowed for high cell density and has low auto-fluorescence. All cultures were grown at 37°C.

5.4.2 Plasmids

All plasmids used have the same backbone containing a kanamycin resistance gene and the P15A replication origin. pTN001 and pTN002 feature a strong PtacI promoter, while pTN003 and pTN004 have a weak Pbla promoter. The pTN001 places the GFP in front of RFP, while all others do the reverse. pTN004 is the only construct with a strong terminator almost directly behind the operon. Sequences can be found in the supplementary data. Harmonisation was performed according to (Claassens et al., 2017).

5.4.3 Fluorescence assay

To ensure an equal growth start, bacteria harbouring different constructs were grown in a pre-culture of 200 µL M9TG, supplemented with 50 mg/L kanamycin in a 96 wells 2 mL Masterblock (Greiner). The Masterblock was covered with a gas-permeable membrane and incubated overnight at 37°C. The pre-cultures were diluted 10000x in fresh 200 µL M9TG and grown in the same way as the pre-cultures. The cultures (and blank medium) were then cooled down to room temperature and diluted 5x in 1x PBS pH 7.4. Finally, 100 µL of the dilution was measured with a BioTek Synergy MX microplate reader at excitation 395/20, emission 508/20, gain 75 (GFP), and excitation 584/9, emission 607/9, gain 100 (RFP). Fluorescence was calculated as raw fluorescence per OD600 for 100 µL 5x dilution. Auto-fluorescence, estimated by introduction of a frameshift in the GFP of pTN002 (pTN002-FS), was subtracted and samples were normalised by dividing by one of the wild-type (WT) samples.

5.4.4 RT-qPCR analysis

10 mL LB with 50 mg/L kanamycin was inoculated 1:1000 from an LB kanamycin preculture. Cells were grown to an OD600 of 0.6 and cooled down on ice-water. Cells were pelleted and resuspended in 250 µL of 50 mM Tris-HCl pH8, 10 mM EDTA and 10 mM DTT. Cells were then lysed with 250 µL of [0.2 M NaOH and 1% SDS]. Protein, genomic DNA and SDS were precipitated by adding 250 µL [1.8 M potassium acetate and 1.2 M acetic acid]. Debris was pelleted in a microcentrifuge tube and 650 µL was transferred to a new Eppendorf tube. RNA was precipitated by adding 650 µL isopropanol and centrifuging for 5 minutes at maximum speed. RNA pellets were washed with 500 µL of [10 mM Tris-HCl pH8 and 70% ethanol], and dried in a laminar flow cabinet. Pellets were dissolved in 100 µL DNAseI buffer (NEB) with 0.25 µL DNAse I (NEB) and incubated at 37 °C for 30 minutes. First, 300 µL of DNAseI buffer was added and then 200 µL of Roti aqua phenol (Roth). The phases were separated by centrifugation and 300 µL of the aqueous phase was transferred to a new Eppendorf tube. 300 µL of isopropanol was added to the aqueous phase and the mixture was loaded on a silica column (Thermo K0702). The RNA was washed twice with 400 µL [10 mM Tris-HCl pH8, 70% ethanol and 100 mM NaCl]. Finally, the RNA was eluted into 50 µL of [1 mM Tris-HCl pH8, 0.1 mM EDTA]. The RNA was diluted to 1 g/L in water and cDNA was generated with the Maxima H minus (Thermo) reverse transcriptase. RT-qPCR was performed with the SsoAdvancedTM Universal SYBR® Green Supermix (Bio-Rad) using cDNA derived from 10 ng of total RNA in a 10 µL reaction.

5.4.5 PSAT region library generation

ssDNA containing 30 degenerate nucleotides flanked on both sides with 4 nucleotide overhang and BsaI recognition sites is converted to double stranded DNA using PCR and a primer that binds in the fixed region. 200 pmol ssDNA and 400 pmol primer is used in a 50 µL OneTaq® (NEB) reaction. 99 Cycles of a 5 second primer binding phase and 5 second elongation phase was performed. The dsDNA is purified and concentrated using a silica column (Zymo D4004). The backbone is prepared by first inserting a substantial piece of nonsense DNA flanked by outward facing BsaI sites between the stop codon and terminator which allows for more precise gel separation later on. The plasmid is sequence verified and pre-digested using BsaI-HF®v2 (NEB) to reduce transformation background. The digested backbone is purified from agarose gel (Zymo D4002). The dsDNA is inserted into the backbone using a NEB® Golden Gate Assembly Kit (BsaI-HF®v2) with a 3:1 ratio. 300 colonies were picked and the fluorescence quantified using a Attune NxT Flow Cytometer (Thermo). 96 cultures covering the full fluorescent range were reinoculated. The cultures were measured again and the associated DNA send for Sanger sequencing.

5.5 Supplementary data

Sequence S5.1: Backbone

```
   1 GCAAGTGGCA CTTTTCGGGG AAATGTGCGC GGAACCCCTA TTTGTTTATT
  51 TTTCTAAATA CATTCAAATA TGTATCCGCT CATGAATTAA TTCTTAGAAA
 101 AACTCATCGA GCATCAAATG AAACTGCAAT TTATTCATAT CAGGATTATC
 151 AATACCATAT TTTTGAAAAA GCCGTTTCTG TAATGAAGGA GAAAACTCAC
 201 CGAGGCAGTT CCATAGGATG GCAAGATCCT GGTATCGGTC TGCGATTCCG
 251 ACTCGTCCAA CATCAATACA ACCTATTAAT TTCCCCTCGT CAAAAATAAG
 301 GTTATCAAGT GAGAAATCAC CATGAGTGAC GACTGAATCC GGTGAGAATG
 351 GCAAAAGTTT ATGCATTTCT TTCCAGACTT GTTCAACAGG CCAGCCATTA
 401 CGCTCGTCAT CAAAATCACT CGCATCAACC AAACCGTTAT TCATTCGTGA
 451 TTGCGCCTGA GCGAGACGAA ATACGCGGTC GCTGTTAAAA GGACAATTAC
 501 AAACAGGAAT CGAATGCAAC CGGCGCAGGA ACACTGCCAG CGCATCAACA
 551 ATATTTTCAC CTGAATCAGG ATATTCTTCT AATACCTGGA ATGCTGTTTT
 601 CCCGGGGATC GCAGTGGTGA GTAACCATGC ATCATCAGGA GTACGGATAA
 651 AATGCTTGAT GGTCGGAAGA GGCATAAATT CCGTCAGCCA GTTTAGTCTG
 701 ACCATCTCAT CTGTAACATC ATTGGCAACG CTACCTTTGC CATGTTTCAG
 751 AAACAACTCT GGCGCATCGG GCTTCCCATA CAATCGATAG ATTGTCGCAC
 801 CTGATTGCCC GACATTATCG CGAGCCCATT TATACCCATA TAAATCAGCA
 851 TCCATGTTGG AATTTAATCG CGGCCTAGAG CAAGACGTTT CCCGTTGAAT
 901 ATGGCTCATA CTCTTCCTTT TTCAATATTA TTGAAGCATT TATCAGGGTT
 951 ATTGTCTCAT GAGCGGATAC ATATTTGAAT GTATTTAGAA AAATAAACAA
1001 ATAGGCTGTC CCTCCTGTTC AGCTACTGAC GGGGTGGTGC GTAACGGCAA
1051 AAGCACCGCC GGACATCAGC GCTAGCGGAG TGTATACTGG CTTACTATGT
1101 TGGCACTGAT GAGGGTGTCA GTGAAGTGCT TCATGTGGCA GGAGAAAAAA
1151 GGCTGCACCG GTGCGTCAGC AGAATATGTG ATACAGGATA TATTCCGCTT
1201 CCTCGCTCAC TGACTCGCTA CGCTCGGTCG TTCGACTGCG GCGAGCGGAA
1251 ATGGCTTACG AACGGGGCGG AGATTTCCTG GAAGATGCCA GGAAGATACT
1301 TAACAGGGAA GTGAGAGGGC CGCGGCAAAG CCGTTTTTCC ATAGGCTCCG
1351 CCCCCCTGAC AAGCATCACG AAATCTGACG CTCAAATCAG TGGTGGCGAA
1401 ACCCGACAGG ACTATAAAGA TACCAGGCGT TTCCCCCTGG CGGCTCCCTC
1451 GTGCGCTCTC CTGTTCCTGC CTTTCGGTTT ACCGGTGTCA TTCCGCTGTT
1501 ATGGCCGCGT TTGTCTCATT CCACGCCTGA CACTCAGTTC CGGGTAGGCA
1551 GTTCGCTCCA AGCTGGACTG TATGCACGAA CCCCCCGTTC AGTCCGACCG
1601 CTGCGCCTTA TCCGGTAACT ATCGTCTTGA GTCCAACCCG GAAAGACATG
1651 CAAAAGCACC ACTGGCAGCA GCCACTGGTA ATTGATTTAG AGGAGTTAGT
1701 CTTGAAGTCA TGCGCCGGTT AAGGCTAAAC TGAAAGGACA AGTTTTGGTG
1751 ACTGCGCTCC TCCAAGCCAG TTACCTCGGT TCAAAGAGTT GGTAGCTCAG
1801 AGAACCTTCG AAAAACCGCC CTGCAAGGCG GTTTTTTCGT TTTCAGAGCA
1851 AGAGATTACG CGCAGACCAA AACGATCTCA AGAAGATCAT CTTATTAATC
```

```
1901 AGATAAAATA TTTCTAGATT TCAGTGCAAT TTATCTCTTC AAATGTAGCA
1951 CCTGAAGTCA GCCCCATACG ATATAAGTTG TAATTCGGTA CCCCGCTTCG
2001 GCGGGGTTTT TTCAAG
```

Sequence S5.2: WT-GFP

```
   1 ATGAGTAAAG GAGAAGAACT TTTCACTGGA GTTGTCCCAA TTCTTGTTGA
  51 ATTAGATGGT GATGTTAATG GGCACAAATT TTCTGTCAGT GGAGAGGGTG
 101 AAGGTGATGC AACATACGGA AAACTTACCC TTAAATTTAT TTGCACTACT
 151 GGAAAACTAC CTGTTCCATG GCCAACACTT GTCACTACTT TCTCTTATGG
 201 TGTTCAATGC TTTTCCCGTT ATCCGGATCA CATGAAACGG CATGACTTTT
 251 TCAAGAGTGC CATGCCCGAA GGTTATGTAC AGGAACGCAC TATATCTTTC
 301 AAAGATGACG GGAACTACAA GACGCGTGCT GAAGTCAAGT TTGAAGGTGA
 351 TACCCTTGTT AATCGTATCG AGTTAAAAGG TATTGATTTT AAAGAAGATG
 401 GAAACATTCT CGGACACAAA CTGGAGTACA ACTATAACTC ACACAATGTA
 451 TACATCACGG CAGACAAACA AAAGAATGGA ATCAAAGCTA ACTTCAAAAT
 501 TCGCCACAAC ATTGAAGATG GATCCGTTCA ACTAGCAGAC CATTATCAAC
 551 AAAATACTCC AATTGGCGAT GGCCCTGTCC TTTTACCAGA CAACCATTAC
 601 CTGTCGACAC AATCTGCCCT TTCGAAAGAT CCCAACGAAA AGCGTGACCA
 651 CATGGTCCTT CTTGAGTTTG TAACTGCTGC TGGGATTACA CATGGCATGG
 701 ATGAGCTCTA CAAATAA
```

Sequence S5.3: H-GFP

```
   1 ATGTCGAAAG GTGAAGAACT GTTTACTGGT GTGGTTCCGA TTCTGGTGGA
  51 ATTGGATGGG GATGTGAATG GGCATAAATT CTCCGTTTCG GGTGAGGGGG
 101 AAGGGGATGC TACCTATGGT AAACTGACTC TGAAATTCAT TTGTACTACT
 151 GGTAAACTAC CGGTGCCGTG GCCGACCCTG GTTACTACTT TTTCCTACGG
 201 GGTGCAGTGT TTCAGCCGTT ATCCGGATCA CATGAAAAGG CACGACTTCT
 251 TTAAGTCGGC TATGCCCGAA GGGTACGTAC AAGAACGTAC TATATCGTTT
 301 AAAGATGACG GGAATTATAA GACCCGAGCA GAAGTTAAGT TCGAAGGGGA
 351 TACTCTGGTG AATCGTATTG AGTTGAAAGG GATTGATTTC AAAGAAGATG
 401 GTAATATTCT GGGTCATAAA TTAGAATATA ATTACAATAG CCATAATGTA
 451 TATATTACCG CTGACAAACA GAAGAATGGT ATTAAAGCTA ATTTTAAAAT
 501 TCGTCATAAT ATTGAAGATG GTTCGGTGCA GCTAGCTGAC CACTACCAGC
 551 AGAATACTCC GATTGGGGAT GGGCCGGTTC TGTTGCCGGA CAATCACTAT
 601 CTATCGACCC AGTCCGCTCT GTCGAAAGAT CCCAATGAAA AGCGTGACCA
 651 TATGGTTCTG CTGGAGTTCG TAACCGCAGC AGGGATTACC CACGGGATGG
 701 ATGAACTATA TAAATAA
```

Sequence S5.4: RFP

```
   1 ATGGCTTCCT CCGAAGACGT TATCAAAGAG TTCATGCGTT TCAAAGTTCG
  51 TATGGAAGGT TCCGTTAACG GTCACGAGTT CGAAATCGAA GGTGAAGGTG
 101 AAGGTCGTCC GTACGAAGGT ACACAGACCG CTAAACTGAA AGTTACCAAA
 151 GGTGGCCCGC TGCCGTTCGC TTGGGACATC CTGTCCCCGC AGTTCCAGTA
 201 CGGTTCCAAA GCTTACGTTA AACACCCGGC TGACATCCCG GACTACCTGA
 251 AACTGTCCTT CCCGGAAGGT TTCAAATGGG AACGTGTTAT GAACTTCGAA
 301 GACGGTGGTG TTGTTACCGT TACCCAGGAC TCCTCCCTGC AAGACGGTGA
 351 GTTCATCTAC AAAGTTAAAC TGCGTGGTAC CAACTTCCCG TCCGACGGTC
 401 CGGTTATGCA GAAAAAAACC ATGGGTTGGG AAGCTTCCAC CGAACGTATG
 451 TACCCGGAAG ACGGTGCTCT GAAAGGTGAA ATCAAAATGC GTCTGAAACT
 501 GAAAGACGGT GGTCACTACG ACGCTGAAGT TAAAACCACC TACATGGCTA
 551 AAAAACCGGT TCAGCTGCCG GGTGCTTACA AAACCGACAT CAAACTGGAC
 601 ATCACCTCCC ACAACGAAGA CTACACCATC GTTGAACAGT ACGAACGTGC
 651 TGAAGGTCGT CACTCCACCG GTGCTTAA
```

SEQUENCE S5.5: LacZ

```
   1 ATGACCATGA TTACGGATTC ACTGGCCGTC GTTTTACAAC GTCGTGACTG
  51 GGAAAACCCT GGCGTTACCC AACTTAATCG CCTTGCAGCA CATCCCCCTT
 101 TCGCCAGCTG GCGTAATAGC GAAGAGGCCC GCACCGATCG CCCTTCCCAA
 151 CAGTTGCGCA GCCTGAATGG CGAATGGCGC TTTGCCTGGT TTCCGGCACC
 201 AGAAGCGGTG CCGGAAAGCT GGCTGGAGTG CGATCTTCCT GAGGCCGATA
 251 CTGTCGTCGT CCCCTCAAAC TGGCAGATGC ACGGTTACGA TGCGCCCATC
 301 TACACCAACG TGACCTATCC CATTACGGTC AATCCGCCGT TTGTTCCCAC
 351 GGAGAATCCG ACGGGTTGTT ACTCGCTCAC ATTTAATGTT GATGAAAGCT
 401 GGCTACAGGA AGGCCAGACG CGAATTATTT TTGATGGCGT TAACTCGGCG
 451 TTTCATCTGT GGTGCAACGG GCGCTGGGTC GGTTACGGCC AGGACAGTCG
 501 TTTGCCGTCT GAATTTGACC TGAGCGCATT TTTACGCGCC GGAGAAAACC
 551 GCCTCGCGGT GATGGTGCTG CGCTGGAGTG ACGGCAGTTA TCTGGAAGAT
 601 CAGGATATGT GGCGGATGAG CGGCATTTTC CGTGACGTCT CGTTGCTGCA
 651 TAAACCGACT ACACAAATCA GCGATTTCCA TGTTGCCACT CGCTTTAATG
 701 ATGATTTCAG CCGCGCTGTA CTGGAGGCTG AAGTTCAGAT GTGCGGCGAG
 751 TTGCGTGACT ACCTACGGGT AACAGTTTCT TTATGGCAGG GTGAAACGCA
 801 GGTCGCCAGC GGCACCGCGC CTTTCGGCGG TGAAATTATC GATGAGCGTG
 851 GTGGTTATGC CGATCGCGTC ACACTACGTC TGAACGTCGA AAACCCGAAA
 901 CTGTGGAGCG CCGAAATCCC GAATCTCTAT CGTGCGGTGG TTGAACTGCA
 951 CACCGCCGAC GGCACGCTGA TTGAAGCAGA AGCCTGCGAT GTCGGTTTCC
1001 GCGAGGTGCG GATTGAAAAT GGTCTGCTGC TGCTGAACGG CAAGCCGTTG
1051 CTGATTCGAG GCGTTAACCG TCACGAGCAT CATCCTCTGC ATGGTCAGGT
1101 CATGGATGAG CAGACGATGG TGCAGGATAT CCTGCTGATG AAGCAGAACA
1151 ACTTTAACGC CGTGCGCTGT TCGCATTATC CGAACCATCC GCTGTGGTAC
1201 ACGCTGTGCG ACCGCTACGG CCTGTATGTG GTGGATGAAG CCAATATTGA
1251 AACCCACGGC ATGGTGCCAA TGAATCGTCT GACCGATGAT CCGCGCTGGC
1301 TACCGGCGAT GAGCGAACGC GTAACGCGAA TGGTGCAGCG CGATCGTAAT
1351 CACCCGAGTG TGATCATCTG GTCGCTGGGG AATGAATCAG GCCACGGCGC
1401 TAATCACGAC GCGCTGTATC GCTGGATCAA ATCTGTCGAT CCTTCCCGCC
```

```
1451 CGGTGCAGTA TGAAGGCGGC GGAGCCGACA CCACGGCCAC CGATATTATT
1501 TGCCCGATGT ACGCGCGCGT GGATGAAGAC CAGCCCTTCC CGGCTGTGCC
1551 GAAATGGTCC ATCAAAAAAT GGCTTTCGCT ACCTGGAGAG ACGCGCCCGC
1601 TGATCCTTTG CGAATACGCC CACGCGATGG GTAACAGTCT TGGCGGTTTC
1651 GCTAAATACT GGCAGGCGTT TCGTCAGTAT CCCCGTTTAC AGGGCGGCTT
1701 CGTCTGGGAC TGGGTGGATC AGTCGCTGAT TAAATATGAT GAAAACGGCA
1751 ACCCGTGGTC GGCTTACGGC GGTGATTTTG GCGATACGCC GAACGATCGC
1801 CAGTTCTGTA TGAACGGTCT GGTCTTTGCC GACCGCACGC CGCATCCAGC
1851 GCTGACGGAA GCAAAACACC AGCAGCAGTT TTTCCAGTTC CGTTTATCCG
1901 GGCAAACCAT CGAAGTGACC AGCGAATACC TGTTCCGTCA TAGCGATAAC
1951 GAGCTCCTGC ACTGGATGGT GGCGCTGGAT GGTAAGCCGC TGGCAAGCGG
2001 TGAAGTGCCT CTGGATGTCG CTCCACAAGG TAAACAGTTG ATTGAACTGC
2051 CTGAACTACC GCAGCCGGAG AGCGCCGGGC AACTCTGGCT CACAGTACGC
2101 GTAGTGCAAC CGAACGCGAC CGCATGGTCA GAAGCCGGGC ACATCAGCGC
2151 CTGGCAGCAG TGGCGTCTGG CGGAAAACCT CAGTGTGACG CTCCCCGCCG
2201 CGTCCCACGC CATCCCGCAT CTGACCACCA GCGAAATGGA TTTTTGCATC
2251 GAGCTGGGTA ATAAGCGTTG GCAATTTAAC CGCCAGTCAG GCTTTCTTTC
2301 ACAGATGTGG ATTGGCGATA AAAAACAACT GCTGACGCCG CTGCGCGATC
2351 AGTTCACCCG TGCACCGCTG GATAACGACA TTGGCGTAAG TGAAGCGACC
2401 CGCATTGACC CTAACGCCTG GGTCGAACGC TGGAAGGCGG CGGGCCATTA
2451 CCAGGCCGAA GCAGCGTTGT TGCAGTGCAC GGCAGATACA CTTGCTGATG
2501 CGGTGCTGAT TACGACCGCT CACGCGTGGC AGCATCAGGG GAAAACCTTA
2551 TTTATCAGCC GGAAAACCTA CCGGATTGAT GGTAGTGGTC AAATGGCGAT
2601 TACCGTTGAT GTTGAAGTGG CGAGCGATAC ACCGCATCCG GCGCGGATTG
2651 GCCTGAACTG CCAGCTGGCG CAGGTAGCAG AGCGGGTAAA CTGGCTCGGA
2701 TTAGGGCCGC AAGAAAACTA TCCCGACCGC CTTACTGCCG CCTGTTTTGA
2751 CCGCTGGGAT CTGCCATTGT CAGACATGTA TACCCCGTAC GTCTTCCCGA
2801 GCGAAAACGG TCTGCGCTGC GGGACGCGCG AATTGAATTA TGGCCCACAC
2851 CAGTGGCGCG GCGACTTCCA GTTCAACATC AGCCGCTACA GTCAACAGCA
2901 ACTGATGGAA ACCAGCCATC GCCATCTGCT GCACGCGGAA GAAGGCACAT
2951 GGCTGAATAT CGACGGTTTC CACATGGGGA TTGGTGGCGA CGACTCCTGG
3001 AGCCCGTCAG TATCGGCGGA ATTCCAGCTG AGCGCCGGTC GCTACCATTA
3051 CCAGTTGGTC TGGTGTCAAA AATAA
```

TABLE S5.1: Construct design

5′UTRs	
5′UTR (tac)	1 TTGACAATTA ATCATCGGCT CGTATAATGT 31 GTGGGGAGAC CACAACGGTT TCCCTCTAGA 61 AATAATTTTG TTTAACTATA AGAAGGAGAT 91 ATACAT
5′UTR (bla)	1 TTCAAATATG TATCCGCTCA TGAGACAATG 31 TGTGGGGAGA CCACAACGGT TTCCCTCTAG 61 AAATAATTTT GTTTAACTAT AAGAAGGAGA 91 TATACAT
IGRs	
IGR [0]	1 ACTAGAAATA ATTTTGTTTA ACTATAAGAA 31 GGAGATATAC AT
IGR [PSAT]	[PSAT] + 1 ACTAGAAATA ATTTTGTTTA ACTATAAGAA 31 GGAGATATAC AT
IGR [PSAT] stem	[PSAT] + 1 CCCCGCTTCG GCGGGGACTA GAAATAATTT 31 TGTTTAACTA TAAGAAGGAG ATATACAT
3′UTRs	
3′UTR No Term	1 ACTAGT
3′UTR [0] Term	1 ACTAGTATAA TGATGTGTTA TCATTGATGC 31 GAGGCGCCTA TACCTCCCCG CTTCGGCGGG 61 GTTTTTTT
3′UTR [PSAT] Term	[PSAT] + 1 CCCCGCTTCG GCGGGGTTTT TTT
3′UTR [16] Term	1 GACCCAACCA CTACACCCCG CTTCGGCGGG 31 GTTTTTTT
PSAT	
PSAT [1]	1 TAATACCAAA CTAAGCTTAA AGAAAGCAAC
PSAT [2]	1 TCGGAAACTA CGTCTTCGTC ATAAACCCTC
PSAT [3]	1 TCCCTCGTTC TACATCTAAT CAACAGCCCT
PSAT [4]	1 ACGATAACGA AACTTTCAAA ACCTAATGAA
PSAT [5]	1 TAAACCACAC TAGCTAGGTA GGAAACAAAA

Table S5.1 Continued: Construct design

PSAT [6]	1 CTCCCCACAT TAGACCTTAG CGGGAACGTC
PSAT [7]	1 TGGTTTCTAC CGAGCGGCCG GCTCCCTCGC
PSAT [8]	1 TTAGCAAACT TAGCCGATTA ATAGAACGAT
PSAT [9]	1 CAATTGAGCA CACCGGACCC AACCACTACA
PSAT [10]	1 CTTTAGTCTC GGTATACTCT TCTGTTTTCG
PSAT [11]	1 CAATTGAGCA CACCGGACCC AACCA
PSAT [12]	1 CAATTGAGCA CACCGGACCC
PSAT [13]	1 CAATTGAGCA CACCG
PSAT [14]	1 GAGCACACCG GACCCAACCA CTACA
PSAT [15]	1 CACCGGACCC AACCACTACA
PSAT [16]	1 GACCCAACCA CTACA
LAYOUT	
pTN001 Series	Backbone - 5′UTR (tac) – GFP – IGR [0] – RFP – 3′UTR No Term
pTN002 Series	Backbone – 5′UTR (tac) – RFP – IGR [0] – GFP – 3′UTR No Term
pTN003 Series	Backbone – 5′UTR (bla) – RFP – IGR [0] – GFP – 3′UTR No Term
pTN004 Series	Backbone – 5′UTR (bla) – RFP – IGR [0] – GFP – 3′UTR [0] Term
pTN005	Backbone – 5′UTR (bla) – GFP – 3′UTR [0] Term
pTN006 Series	Backbone – 5′UTR (bla) – GFP/RFP/LacZ – 3′UTR [PSAT] Term
pTN007 Series	Backbone – 5′UTR (bla) – [see Figure 5.6] – 3′UTR [PSAT] Term

TABLE S5.2: PSAT library sequences with observed fluorescence

PSAT sequence (GFP lib)	GFP	PSAT sequence (RFP lib)	RFP
CTTTAGTCTCGGTATACTCTTCTGTTTTCG	38257	TAAACCACACTAGCTAGGTAGGAAACAAAA	98994
AGTTGTCCGTGCGTCTCTTAAACTGGTAAA	34595	GCGAAGTCCAACACTCCACCAAGAATCTAC	90809
CAAGTAAGCTTGAGGCCTAACTACAACAAA	34401	TCAAACACAATTCAATCTACAGCAAACAAG	89434
ACCTTACTTCGCTTAAAACTCTGTATCTAA	33026	AAGAGAGCAGTAGAGGCGTCAGCAGGACAC	85051
CGAACGGGTTAGACGTATATAAACTGAAAA	32869	TAACCAAGCAAAGAAACCACATCCCACTAA	84717
GACCTCGCCCACCCAAGTTGCACCTATGTA	32795	TAACGACAATCACGGTGTGAGAAATCTATC	82093
CAATTGAGCACACCGGACCCAACCACTACA	32143	TCTGGCATTCCTCCCGGCGTGCTACCTCAT	81512
TCCATGTCCCGCACCTTTCCCTATTCTACT	31964	ACGATAACGAAACTTTCAAAACCTAATGAA	77448
CCTGATCAAATTATGAAATAAACCTCTGAA	31957	TAAAACCGCGAAAGCATCACAACAAACCAA	76806
CCTGTAAAGCGAACGAGCAAAACTCATACA	31923	GGAGGTAAAGATAGTCAAACACAACAAGAA	75679
CGCCACCAGACATGCCCGTTCTTACTAACC	31792	ACAAAACTCAGAGGAAAAGAAGAAAACAAA	74712
CCCTCCACTTAATGGCAAGCAGTCCTTCCA	31775	GGAAACCGCAGACGATAAATAGGAGCAAAA	74281
ACCGGGCCCCACCCCGTTTGCAATACCTCA	31770	AGGACGTTCGAAGGTAACAATATGAGAAAT	73571
CTCATTTGCTACTCACTTTATAGCACTGTA	31694	AGTGGCAGCTCAGCATCCTTTGTACCCTAA	72939
TCTTGAATGTTAAGCATGCAGTTAATACAG	31690	AATGGCACAACATCCAAAATCTAAAACCAC	72748
AGCCTGACCTTGATTATGAGAGTGAACAAA	31668	CCGACACGAATGGCCGACCGAAGTAATACA	72500
CTTCACGTTCACGCCTTTACAATTGATTTA	31649	TCCTAAGATCACCTTTCCATCCTAACCGAC	71527
CCCCTATCATCGCTTCTGGTACATCACCTA	31542	TTACCAAAATCCAACTCAACAAAGAAATAT	70909
AGGGCCTTCCCTTCCCTACCTCTGTCCAAC	31467	AACACATTATCTCACTTTTAATCACGTTAA	70063
ATCTCATTCGTGATTGTATATGATTGAGTC	31345	AAAACCGCACAAATATCCAATAGGCGCAAA	70046
AGGTTCGTTGGTGTTTATAACACTTTGTTT	31229	ATAAGAAATAAAACAAAGTAAGTAAGATCA	70022
CTCAGAGCCCCTGTAGTCAGCACTTTGCCC	31106	TCGCACGGCTCAATGTGCAAATTACACCCA	69908
ACCACACGATTAAATCCCAGAAACATCATA	31043	CAAATGTTCGCGCGCAGTGCGCGTTGGTAG	69856
ACTAAGCTTTACATAAGGCTGATTGTGCAC	30955	GCACATATGAGCAGTACGAGAGCAATATAA	69460
ACACGTTTCCGGTGTTGGCCCATCACGATA	30935	CAAAACAAAGGACACGCCAAAATAATTTAC	69409
ACATGAGCGGAATCGCTAACTAAGTTAAAC	30902	TAACTCTCAAGACGCGATACCAAAACATAA	68868
AATTGTCGATGTATGCTAAAACTTCAAATT	30835	ATAACCTGACATACCCCTAAGATAACCGTG	68078
CTCAATCCATAGACCAATCCAACCAATTCT	30833	GACACCTCCCAAACCACTGCACCTTGAACC	68042
ATCGATATTCCGCTAGATATATGTTCATTT	30812	AAGCAATACACAGATAATAAACACACAAAT	68006
CTGGGCGTCCAACTAAGGCCCCACGGACCT	30601	AAGAAATAAACCATAACCAAAATCAGTGTG	67633
ATCGACACCCTACCGGACAACTTTGTCTGC	30495	ACGGCTCTTTGAGACGTCATGTTATACTCC	64287
TCCTTTCTGCAGTAAGAAGTAAACGAGAAT	30487	TGGGCCTGGAGCGCCACCGTACCGGAGGAG	63889
TCCATGTCCCGCACCTTTCCCTATTCTACT	30410	CCGTTATGATATCCCTCTTAAACATTCTAC	63853
CTCAGAGCCCCTGTAGTCAGCACTTTGCCC	30372	AACATGGCTACGGATCCAATGCCACATGTT	62205
TAGCTAGGTTGCTTGACAATCTGTCCCTTC	30331	TCCCTCGTTCTACATCTAATCAACAGCCCT	57534
AATTTTAGGCTATTACGAAGACTTGTTATT	30163	ACATGCTGAGTTTTCGAATCGGATCGAAAA	54974
CTGAAATCACTAATGTTTCGGTAAAACGCT	29730	TACCCTTCTTCAGCTGCTTTCCAACCCTCC	52651

Table S5.2 Continued: PSAT library sequences with observed fluorescence

PSAT sequence (GFP lib)	GFP	PSAT sequence (RFP lib)	RFP
TCCTGCCGCGCGCACGGTTGGGCAACGGCA	29702	ACAGAATGCGTGGGCGCGGAAGGAGCAAGC	50340
TCCAGGCAAAGGCACCCTTCGAAACGCACT	29652	CCATGAATCGTCTGCGTCGCTGGGGCTCTC	50039
ACTAGTTCTGGTATCATTAGGTCTAGTTGC	29584	CCTCTCCTATATTCCTCCACCGCATGCTAT	49081
AATCTTCCTGCACTCACTAGCCGTCTTATT	29553	TAATGAGGTCACGGTGTGCTGGAAGGGTGT	49015
CTCACAGACCTTCCTCACCCATCTGACTCC	29152	ACGAACCTTTCCACTCCCCAATTCTTAAGT	48558
CTGAAATCACTAATGTTTCGGTAAAACGCT	29076	ATGAAAACTACAAAATTCATCAACTAAACT	48462
TCGCCTTCAACAGGGCCTATCCAGTACCCC	28840	ATCCTGCCCTTCAGCCCATGCTCCTCCTGT	47859
ATCAGACACCTTATGACTACCAGTAAAGTC	28709	CCTAGCTGGGAGCGCGCGGTGTGCGTTGCC	47590
CTTGTTGAGTGGTGCCTCGGGGAGCGAGGG	28626	AGTGAAATGCTAGCTACGCCGTTCTCCTTC	46727
TGCCTGCGCGGCGCCCCGGGCGACAGGCCC	28293	TCCGATCCTTTGGACTCCGCGGCGGCTCGT	45352
CCACGGTGAAGCTAATCACCTCCCGGTGCC	28148	ATTCTCTCACCTCGCACCCAGGCGGGCGCC	44284
CCAAAGCCATGGCTGTCTGCAACCAAAGAC	27866	TAGTACAAATTTGCTAACATAAATACAATC	42824
CCGCTCTTCAGAGCGCCTAATCTCCGAGCC	27341	ACGAACCTTTCCACTCCCCAATTCTTAAGT	41525
ATCAATGGTTAGCGTAATCGACACTATACC	27157	TCTTCCTACTCTCTCCTACTGTCTCCTTTC	41331
TTCTTTGCGTTCTAACATGTTGCGTATACT	26959	TCGGAAACTACGTCTTCGTCATAAACCCTC	39852
CATCACAGACCCAAGAGCCGCAAAATCGTC	26574	TAACTAATACATACCACTGAGATCTCCTCC	36239
TCATACCTAGCTCCTAGTACATTCCCCGCG	26496	TAGCACACAAAACGAAATAAATCAACCAAG	35361
TTAGCAAACTTAGCCGATTAATAGAACGAT	26413	AATATACCAAGCAAATACAGGTGACGCAAT	30280
ACCCGAGAACCTCCCCGCCCGCTCCCCACT	22798	TAATACCAAACTAAGCTTAAAGAAAGCAAC	18374
TCGAGGAGGGTGTGGCGAAATCTCAGTGCT	22104		
CGATTAAACGCCATAGCAGCGTGGGGCGGC	22018		
CGGGACCCACAGGGCCCAGTACCCTGTGGG	21922		
CGGGCGTAAGGGCGCGTGCGGCGGTGTGTG	21553		
ACAACAGAGTCCGACCGAGAGGGGCCGAGT	21261		
CCAGTTGTGCGTCACCTTGTCATTGTTTGT	21059		
ACTACAGCTATGCTCCGAATCTACAGGAAA	20677		
TCTTTCGGCTCGTGGCGCGCGCTCCCCGCG	20622		
TGGTTTCTACCGAGCGGCCGGCTCCCTCGC	19503		
TCGTGCTCGGCATATGCGGGCGGGGCAATA	18815		
GCGTGTTCCTATTTCCATTCATGTAGGTAT	17980		
CCGGCGGGTGTGCGCGCGTGTTCCGCCGCT	16654		
TCGCAGAGCTGTATTAAGCCCATTGCAATC	16098		
CACCATCCACACGTCGAAGCATATGTTAAT	15087		
CGCGCGTGCCAGTGTGGGTGGGCGGGCGGC	13991		
CTCCCCACATTAGACCTTAGCGGGAACGTC	13722		

Chapter 6

Summary and general discussion

6.1 English summary

Protein production in cells can be influenced by many regulatory elements both at the DNA and at the RNA level. A cell has many control steps that influence gene expression through tuning both transcription and translation rates in order to produce the right protein, in the right amount, at the right time. These regulatory elements can be modified to force the cell to optimally produce any protein of interest. This thesis focuses on translational regulatory elements. These elements can be categorized into the three parts that make up a gene: the 5′ untranslated region, the coding sequence and the 3′ untranslated region. These regions were studied to better understand their regulatory mechanisms and thereby enable tuning of these regions.

First of all, the relevance of protein production, both in an industrial and academic setting is discussed in **Chapter 1**. An overview is given of the transcriptional and translational features that contribute to overall protein production and the main problem is put forward that optimizing these features towards high protein production is complex. The reason for this is that many of these features are interconnected, meaning that if the sequence of a genetic element is changed to optimize that element, the performance of other elements can be influenced simultaneously, which in the best case may result in the desired stimulating effect, but in the worst case may lead to complete inhibition of an essential process. This complexity is the reason why in this work the focus has not only been on fundamental discoveries but also on the development of a generic practical approach to optimize functional protein production. Finally, an overview is provided of distinct, frequently-used protein production systems with their pros and cons.

Chapter 2 provides an overview of many recent studies that have been conducted in the quest to optimize protein production. Advances in big data generation and analysis contribute to a more thorough understanding of many of the factors involved. Particularly in eukaryotes, a strong link has been found between translation elongation rates and mRNA stability. A transcript with a fast translation rate is more stable as, presumably, the high density of fast-moving ribosomes protects the transcript from a degradation process initiated by the binding of RNA-degrading enzymes. Furthermore, studies focusing on translation initiation are using new tools that can experimentally determine RNA secondary structures *in vivo* as opposed to the currently much used *in silico* predictions. These developments will contribute towards understanding the impact (good and bad) of secondary structures in the translation elongation process. An overall conclusion is that the effects of less influential genetic features are difficult to distinguish as these effects are often overshadowed by other features that may be stronger and may change simultaneously. A prime example of this is the effect of codon usage on translation elongation, which is mainly hidden by changes in translation initiation induced by changes in secondary structures upon codon alterations. Machine learning approaches may offer a solution for revealing these more nuanced, subtle effects. The good news is that this may result in better predictability of sequence features. The bad news may be that, due to their "black box" nature, these approaches may not lead to increased biological understanding.

In **Chapter 3** a genetic design was studied that can decrease the effects of secondary structure on translation initiation, which is a main limiting factor in the overall translational process. A bicistronic design is a naturally occurring element in which the open reading frames of two genes overlap. This overlap probably causes a reduction in secondary mRNA structures involving and surrounding the ribosome binding site of the second cistron, i.e. the gene of interest. The explanation of this phenomenon is that the ribosome translating the upstream gene exhibits helicase activity throughout the translation initiation region of the downstream gene. A tool like this allows for studying the relatively subtle contribution of codon usage (Chapter 4), because effects of secondary structures involving the ribosome binding site often overshadow the effects of codon usage on translation elongation (as discussed in Chapter 1 and 2). To test the potential of this design, 11 codon sequences were expressed with and without a bicistronic design. We observed that a bicistronic design can drastically improve the expression level and that it changes the relative performance of different codon optimization strategies. This was expected as reducing the impact of translation initiation limitations will reveal the impact of only translation elongation (including codon usage). We further showed that the bicistronic design has the potential to completely rescue the expression of constructs that were limited by a strong secondary structure including the ribosome binding site. We concluded that the incorporation of a bicistronic design is highly valuable in general expression vectors as they consistently improve protein production levels. Finally, when benchmarking different codon optimization algorithms or analysing codon usage, the use of a bicistronic design will amplify the effect of codon usage on translation elongation. This will give a more fair comparison as the effects are less dependent on the 5′UTR sequence.

In **Chapter 4** the effect of codon usage on overall protein production was explored. As discussed in Chapter 3, a bicistronic design is invaluable in codon usage studies when the effects of translation elongation are targeted. We, therefore, included a bicistronic design in all our generated codon sequences. To observe the effects of codons on translation we opted for synonymous codon randomization of the complete coding sequence of the gene as opposed to codon randomization within a specific region as was often done in previous studies. We generated a huge number of different codon sequences (350.000) all encoding the monomeric red fluorescent protein (mRFP). A subset of 1459 variant colonies was selected covering the full range in expression. The good-performing sequences were found to outcompete commonly used (commercial) and recently proposed (academic) codon optimization strategies. This indicates the potential for the developed randomization approach, which would be a generic method for optimizing protein production if it can be coupled to high-throughput screening/selection (see discussion). Two different machine learning algorithms were trained on our data set, and for our best performing algorithm, we obtained a Pearson correlation of 0.803. We further used a sliding window approach to feed the algorithm limited information and observe at what position the highest level of predictability is reached. In essence, this approach allows for detecting the limiting factor in the translational process. Somewhat to our surprise, we saw that despite the use of a bicistronic design the majority of expression differences could still be explained by the codon usage of the first 9 amino acids, and by the secondary structure formations including the first 9 codons and the 5'UTR. Codon usage throughout the remainder of the open reading frame did influence the translation efficiency but to a lesser extent. We can now conclude, due to our full randomization approach, that secondary structures around the RBS and not the overall codons usage is the primary determinant of translation efficiency, and hence of protein production.

In **Chapter 5** we discovered a new type of translational coupling. We observed that in an operon design, the translation efficiency of a downstream gene can influence the translation of an upstream gene. This coupling may have relevant implications when making changes (insertions, deletions, rearrangements) within operons. That also means that phenotypical effects cannot be exclusively assigned to a gene knockout within an operon, as that gene itself may also influence the expression of up and downstream genes. Furthermore, we discovered the substantial effect that a transcriptional terminator can have on the translational process (50-fold increase). The 3'UTR was examined in more detail by inserting a randomized 30 nucleotide sequence between a CDS' stop codon and the terminator. This region was found to have the potential to influence the overall protein production, adding yet another tunable region to the toolbox. Interestingly, unlike 5'UTR sequences, this sequence acted independently of the genomic context. A good-performing 3'UTR sequence leads to high expression regardless of the CDS they are placed behind. The intriguing possibility exists that this may be a generic control system (at least in bacteria), although this has to be demonstrated. What we do know at this stage is that this region is useful if reliable fine-tuning of gene expression in *E. coli* is desired.

6.2 Discussion

The production of proteins in prokaryotic and eukaryotic microorganisms as well as in mammalian cell lines is not a trivial task. Despite the vast amount of knowledge that accumulated over the past 50 years, regarding the genetic code and the various elements that influence protein yield and folding, there are still no fixed recipes for guaranteed, successful protein production. This limits fundamental work, for instance when sufficient protein needs to be obtained for structural and functional studies, but also industrial applications where some processes are not economically feasible or very costly due to limited production.

Studying an individual genetic element in a fixed genetic context can elucidate its fundamental mechanism and even lead to optimization strategies for that element. For instance, a large amount of RBS sequences and secondary structures involving the 5′UTR of the downstream gene were generated and measured to develop the RBS calculator (Salis, Mirsky, and Voigt, 2009). This tool allows for predicting the translation initiation rate. The predictions are based on mRNA secondary structure predictions, and on interactions between the 16S rRNA and the ribosome binding site. The model was verified by expressing a myriad of RBS sequences in combination with two chimeric RFP proteins containing the first 27 nucleotides of either TetR or AraC. The model was able to predict the expression levels of chimeric fusions reasonably well (TetR-RFP R^2 0.54 and AraC-RFP R^2 0.95). However, as seen in Chapter 3 (Figure S3.3), RBS calculator correlations were weak for our mRFP expression data. Algorithms like the RBS calculator can be used for approximations of translation rate, but they cannot account for features such as gene dosage or promoter activity (Jeschek, Gerngross, and Panke, 2016). This example shows that predicting the efficiency of genetic elements in a controlled context has merit, but predictions become less reliable as soon as more genetic variations are introduced. This is the reason why optimizing a full gene for a particular production level in a single attempt has a low chance of success.

The interactions that exist between the many gene features are complex and difficult to predict with current tools. Therefore, if a particular level of transcription and translation is required, a combination of universal genetic elements and randomization and selection is needed. To start, the transcriptional rate should be low to first optimize for translation without burdening the cell. Then the translational process can be optimized by changing the 5′UTR, the CDS and the 3′UTR. Ideally, genetic elements are used that reduce the inhibitory effects or act on translation in an independent manner. Examples of this have been described in this thesis with the exploration of the BCD in Chapter 3 and the discovery of a universal tuning sequence in the 3′UTR of prokaryotic genes in Chapter 5. These steps will likely already improve gene expression to a sufficient rate. However, in some cases, or if the production should be further increased, the codon usage also needs to be altered to optimize the overall translational process. To date, there is no sure-fire way to achieve this with prediction tools. Randomizing the codons in the 5′ part of the CDS followed by *in vivo* selection for optimal gene expression appears to be a good way to ensure a good performing codon sequence (see Chapter 4 and Discussion 6.2.4). After translation optimization, the transcriptional rate can be increased in a step-wise manner until the desired production rates are reached.

6.2.1 Optimizing the 5′UTR

The use of a BCD showed great potential for the improvement and normalization of the translation initiation efficiency in prokaryotes. This naturally occurring overlap of genes has the potential benefit of translational coupling (Huber et al., 2019). By coupling the translation activity of a fixed short peptide to the POI, a more constant level of translation may be expected. This effect likely takes place due to the RNA helicase activity of the ribosome and the translation termination and reinitiating effect that can take place at the overlapping stop and start codon of the two CDSs (Huber et al., 2019; Mutalik et al., 2013a). The potential of this design as a standardized module for gene expression in *E. coli* has been tested in Chapter 3, as we observed a consistent increase in gene expression upon utilizing a bicistronic design compared to a monocistronic design. More specifically, for one of the coding sequences tested with a monocistronic design, a secondary structure involving the RBS seemed to completely abolish the expression by blocking the ribosome from binding. Its expression was substantially recovered by the incorporation of the BCD, close to the level of a rational design to remove the inhibitory secondary structure (Figure 3.2). In this specific case, it was obvious that the RBS was not accessible so a rational design to circumvent this inhibition was possible. Although this was not the case for the other constructs, the use of a BCD still more than doubled the expression levels. This highlights the limited predictive power of secondary structure algorithms to identify more subtle interactions. Modules like the BCD offer great potential in standardized expression vectors as they can improve expression without requiring any rational design by the user.

An interesting hypothetical addition to the bicistronic design would be to increase the ribosomal throughput by designing a synthetic looping of the 3′UTR back to the 5′UTR. Incorporating complementary sequences in the 3′UTR to the 5′UTR in such a way that the stop codon is in close proximity to the ribosomal

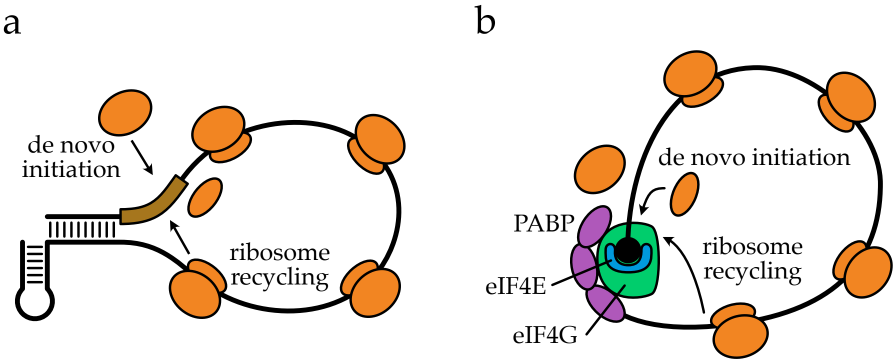


FIGURE 6.1: **Closed-loop structures in prokaryote and eukaryote.** (**a**) A hypothetical prokaryotic closed-loop structure. Complementary sequences between the 5′UTR and 3′UTR bring the stop codon and translation initiation site (depicted in brown, can be a monocistronic or bicistronic design) close together allowing ribosome recycling. (**b**) The eukaryotic closed loop structure promotes ribosome recycling via interactions between the 5′cap and poly(A) tail due to interactions between poly(A)-binding protein (PABP) and eIF4G.

binding site could increase ribosomal density at the translation initiation site (Figure 6.1a). This idea is based on the eukaryotic expression mechanisms, where the 5′ cap-interacting factors and the 3′ poly(A)binding protein interact and form a closed-loop structure promoting both translation and mRNA stability (Thompson and Gilbert, 2017) (Figure 6.1b). The complementarity between the 5′ and 3′UTR could also enhance mRNA stability, on top of the protective role of the terminator structure, by acting as a second line of defence against the RNAse R and PNPase; these ribonucleases can only digest single-stranded RNA (Guarneros and Portier, 1990; Mohanty and Kushner, 2003). However, the double-stranded RNA section should not be too long as it otherwise could invoke degradation by RNAse III (Nicholson, 2014) and the complementary section should not be too close to the open reading frame as it could sterically hinder the ribosome from reaching the stop codon. An ideal length appears to be 30 nucleotides based on a study by Li *et al.*, where they tested different lengths between the stop codon and terminator and found that inhibitory effects occur when the spacing is less than 30 nucleotides (Li et al., 2016).

Eukaryotic translation initiation differs from prokaryotic translation initiation as it does not use a ribosome binding site (see Chapter 2). Therefore, it is not possible to apply a BCD in the same way to eukaryotes. However, there are other ways to standardize the 5′UTR in eukaryotes. A commonly used, perhaps somewhat inadvertent way to reduce secondary structures in the 5′UTR is the use of N-terminal purification tags such as a polyhistidine tag or a strep-tag. These tags are used to purify the expressed protein via affinity purification. They can also be seen as standardized sequences that reduce the chance of inhibitory secondary structures in the 5′UTR region or enhance solubility (Ki and Pack, 2020). If a multitude of proteins is produced with the same N-terminal purification tag and expression construct, optimization of the codon usage within the purification tag will be beneficial. It should be noted however that an N-terminal purification tag cannot always be used as in some cases it influences the protein folding and results in inactive protein. Also, for therapeutic applications, non-native features are generally not desired. A second approach in eukaryotes to perhaps achieve an effect similar to the BCD in prokaryotes would be to incorporate a short peptide with a 2A self-cleavage peptide, which was first discovered in the food-and-mouth disease virus (Ryan, King, and Thomas, 1991). By using the 5′ of a well-expressed gene, translation initiation rates could be improved. A 2A sequence could then split the POI from the peptide (leaving 1 non-native amino acid). These two approaches which standardize the 5′UTR and 5′CDS are a good alternative for more consistent gene expression results until design principles are fully known and can be reliably exploited.

6.2.2 Optimizing the 3′UTR

The BCD standardized the 5′UTR efficiency and can therefore also be used more reliably as a tuning tool (Claassens et al., 2019). We discovered that the 3′UTR, more specifically the non-coding sequence between the stop codon and intrinsic terminator sequence, also shows tuning potential. The influential effects of this 3′UTR sequence appear to be independent of the CDS in question. Interestingly, a supportive observation was made in Chapter 4 (Figure 4.5c) where a clear dip in the 3′UTR is observed suggesting that this region holds no information regarding the expression changes due to codon alterations. This further substantiates that although this region does affect overall protein production, it is not due to interactions with the codons. Although we cannot yet explain the mechanism behind this phenomenon, we hypothesize that this sequence either stabilizes or destabilizes the rho-independent terminator possibly through secondary structure interactions. The stem-loop length and GC ratio of a rho-independent terminator define the stability of the terminator. One of its functions is to prevent exoribonucleases from digesting the transcript as RNAse R and PNPase cannot digest double-stranded RNA (Mohanty and Kushner, 2003; Mohanty and Kushner, 2016). Another hypothesis is that the bases downstream of the stop codon interact with the ribosome and aid its release, possibly preventing a ribosome traffic jam near the end of the CDS (Tate, Cridge, and Brown, 2018). Finally, more complex, long-distance secondary structure interactions with the 5'UTR could take place protecting certain regions by forming double-stranded RNA, or feeding ribosomes back to the translation initiation site as was also suggested in a design in Figure 6.1a. Despite the lack of an explanation, this region can be a very useful tool for additional turning of expression, and it is a good candidate for optimization in expression platforms.

A stabilizing effect of bases in the 3'UTR has also been shown in eukaryotes. Poly(U) sequences act as stabilizing elements by interacting with the poly(A) tails, preventing the binding of poly(A) binding proteins. This disrupts the natural regulatory function of poly(A) tails and the stem-loop that is formed between the poly(A) and poly(U) sequence stabilizes the mRNA, seemingly analogous to the rho-independent terminator in prokaryotes. These 3'UTR poly(U) regions were observed in 10% of *S. cerevisiae* genes and appear to be sequence context-independent, similar to the independent effect of the 3'UTR in prokaryotes. Transplanting poly(U) sequences into different genes showed increased mRNA stability for 5 out of 6 tested cases (Geisberg et al., 2014). Altogether, it is clear that the 3'UTR shows potential for tuning mRNA stability and expression. Surprising but useful is that the effects appear to be sequence-independent making this a reliable protein production tuning region.

6.2.3 Optimizing the codon usage

While advances have been made regarding the *a priori* codon optimisation designs, it is still not possible to generate a codon sequence for high protein production with a high degree of reliability. It has become clear from other studies and the data presented in Chapter 4 that the majority of the effects on translation originate from the 5'UTR and codon usage in the 5' end of the CDS (Cambray, Guimaraes, and Arkin, 2018; Kudla et al., 2009).

The codon usage through the remainder of the CDS also influences expression levels as shown in Figure 4.5c and 4.6c. However, the simple notion that common codons (instead of rare codons) are required for good translation, which still lies at the basis of many optimisation algorithms, is not that clear. In our dataset (Chapter 4) we observed constructs that contained many rare codons (mainly in the CAI_M and CAI_L set), for many of which *E. coli* does not possess a matching tRNA, but are still leading to translation levels similar to that of constructs containing only common codons (Figure 4.3 and 4.4). We hoped to discover the rules of these more subtle codon interactions by using a bicistronic design, which would reduce the known overruling factor of 5′UTR interactions and amplify the more subtle codon interactions. However, despite this precaution, we still found the interaction of codons with the 5′UTR as the main limiting factor. This does, however, highlight that the focus of codon optimisation for prokaryotic gene expression probably should be on interaction with the 5′UTR rather than the usage of common codons.

The use of machine learning in the analysis of codon usage data shows great potential (Pearson correlation: 0.803, Figure 4.8). However, in order to generate a more universal predictive algorithm a lot more data is needed, particularly codon data from other proteins. Because codon variation in our design occurs only at the third nucleotide position, the first and second nucleotides remain constant throughout the dataset. This means that the algorithm, when using our all-inclusive one-hot featurization, cannot learn what the effect would be when these nucleotides would change, for instance when another amino acid sequence is used. Similarly, the algorithm is trained on a constant 5′UTR so it cannot predict what will happen when a different 5′UTR is used. The more general base pair probability (BPP) featurization, which is based on *in silico* predicted secondary structures, is more universal. The BPP feature can be calculated for new combinations of 5′UTR and CDS but also when amino acids change. However, the predictive power of this feature is weaker but still substantial when using a random forest regressor and looking at the 5′CDS (Pearson correlation of ~0.6, Figure 4.6). This likely originates from limitations of *in silico* secondary structure prediction algorithms, particularly for weaker structures or long-distance interactions, and might improve over time as these algorithms improve or because secondary structure alone cannot explain the overall protein production rate.

6.2.4 Screening and selection as an alternative to rational design

To apply gene optimization in an industrial setting, optimization predictions need to be reliable. Since this is still not the case for codon optimization strategies an alternative would be a quick and efficient screening approach. Because the costs of synthetic DNA is rapidly decreasing, it becomes affordable to generate a huge number of gene variations and select the best performing ones (Carlson, 2009). However, the selection of well-producing variants is not straightforward when the protein cannot be directly quantified in a high throughput manner, which is only possible for specific proteins such as a fluorescent protein (Figure 6.2a). Manual sampling-and-measuring of expression from many constructs is inefficient and quantification methods like SDS-PAGE are not able to distinguish the subtle differences in expression. In order to indirectly quantify the expression of a protein, there are several genetics constructs available that couple the level of translation to a reporter gene (Figure 6.2b). A direct fusion of the protein of interest (POI) to a reporter protein will result in a perfect 1:1 expression ratio (Figure 6.2b-I). However, a fusion might impair the POI function due to protein misfolding or more importantly, might influence the reporter signal strength. Additionally, the mRNA can form drastic new secondary structures upon the addition of the reporter CDS. These structures could influence the expression of the POI. Or perhaps, more importantly, efficient sequences that are identified using a translational coupling system might be dependent on the translation of the reporter sequence. Ideally, after initial screening, the reporter can be removed from the construct in an industrial setting to preserve cellular resources and increase the production of the POI. However, important secondary structures, or interactions with the 3′UTR, might be disturbed when the reporter is removed. An alternative would be to disable the coupling system and stop the translation of the reporter. E.g., a stop codon could be introduced which results in the loss of reporter translation but with minimal changes in the mRNA sequence. However, as observed in Chapter 5 the translation of a gene is also influenced by the translation activity on the surrounding mRNA due to translational coupling. Thus, even disabling the coupling but leaving the reporter sequence in the transcript will have some effect on the overall levels of the POI production.

Besides a fusion of the POI and a reporter protein, there are coupling systems that result in separate proteins. These systems do not suffer from potential protein misfolding as the proteins are no longer fused. A naturally occurring translational coupling system is the overlap of a stop codon and a start codon (e.g., **ATG**A, TG**ATG**, or TA**ATG**) also known as "termination-reinitiation" (TeRe, Figure 6.2b-II, Huber et al., 2019). When the ribosome reaches the stop codon of the first ORF it dissociates but can reinitiate at a methionine codon in close proximity. The rate of reinitiation is relatively low and translation of the second gene is influenced by the presence of RBS like sequences in the 3′UTR of the upstream gene. However, TeRe also takes place in the absence of an intragenic RBS (Huber et al., 2019). The use of a TeRe system for the identification of highly expressed genes has potential especially because the ratio between POI and reporter is not 1:1 but likely much lower. This means that fewer cellular resources are spent to synthesize the reporter, allowing for a wider identification range of the POI. A downside of this system is that a codon variant of the POI that has an RBS-like-sequence in its 3′ CDS will result in higher production of the reporter, which is not correlated to the overall expression rate of the POI. This means that selection could favour perfect RBS sequences in the 3′UTR of the first gene, instead of a high translation rate of the POI. A semi solution to this would be to exclude the

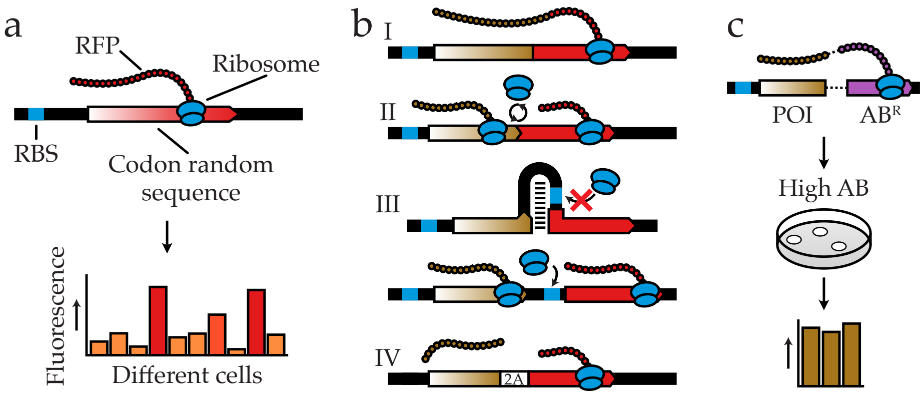


FIGURE 6.2: **Translational coupling.** (**a**) A codon randomized RFP gene (red gradient) will result in many different levels of expression when expressed in a cell. The selection for the best expressing variant can be done directly with FACS. (**b**) Four translational coupling systems correlate the expression of the codon variable protein of interest (POI, brown gradient) to the expression of a fluorescent gene (red). **I**; A direct fusion of the POI to the reporter protein. **II**; An overlap of the stop codon of the POI with the start codon of the reporter (ATGA, TGATG, or TAATG). **III**; The TARSyn system, which blocks translation initiation of the reporter due to secondary structure formation with the RBS, only upon translation of the POI and due to the helicase activity of the ribosome the RBS of the reporter becomes accessible. **IV**; A fusion of the POI and reporter protein with a 2A self-cleaving peptide which separates the protein during translation. System III works only in prokaryotes and system IV works only in eukaryotes. (**c**) An antibiotic resistance marker can be used to select for high expression based on viability.

3′ of the POI from randomization. Since the majority of the changes in protein production originate from codon 2 to 9 (Chapter 4) this will likely still result in a wide range of expression.

A third coupling system, which is exclusive to prokaryotes, is the TARSyn system (Rennig et al., 2018, Figure 6.2b-III). The translational coupling of the gene of interest and reporter is based on the inaccessibility of the reporter's RBS due to a strong secondary structure. Strong secondary structures around the RBS region can severely hamper translation initiation (Boël et al., 2016; Cambray, Guimaraes, and Arkin, 2018; Goodman, Church, and Kosuri, 2013; Kudla et al., 2009) (Chapter 3 & 4). Due to the translation activity on the gene of interest and the helicase activity of the ribosome, the reporter's RBS becomes available every time a ribosome is active on the 3′ of the POI's CDS. Similarly, to the TeRe system, the ratio between the POI and reporter will not be 1:1 but lower for the reporter. Also, the TARSyn system can potentially be influenced by codon usage in the 3′ of the POI, strengthening or weakening the intended TARSyn secondary structure. In the intended design, only the stop codon of the POI and the first two codons of the reporter are part of the stem structure, however, if the codon preceding the stop codon also interacts with the third codon of the reporter the stem-loop would be strengthened. This could influence the coupling and lead to false conclusions regarding efficiently expressed codon sequences.

The fourth coupling system relies on the aforementioned 2A self-cleaving peptide (Figure 6.2b-IV). During the synthesis of this peptide, the ribosome makes an error and skips a peptide bond causing the synthesized protein to be split. 2A self-cleaving peptides have only been found in eukaryotes and do not work in prokaryotes (Ryan et al., 1997). They do however allow the linking of the POI to a reporter with a 1:1 ratio but do not suffer from potential protein misfolding, unlike direct fusions.

We have performed some preliminary experiments to assess the potential of these kinds of selection systems for selecting optimal codon sequences from a library. To see if these reporter systems can properly link the expression of the POI to the reporter and do not influence the relative expression level of the POI, we used 5 codon sequences that range from low to high mRFP expression (mRFP-1 to mRFP-5 from low to high expression, selected from the codon random library generated in Chapter 4) and attached a GFPuv reporter sequence for each coupling construct. The variable mRFP acts as the POI that is to be quantified using the GFPuv as a reporter.

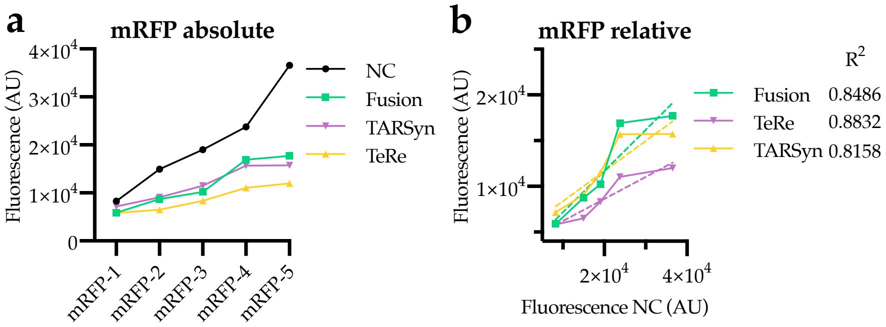


FIGURE 6.3: **Influence of coupling system on mRFP expression.** (**a**) Absolute mRFP fluorescence of codon sequence mRFP-1 to mRFP-5 without a coupling system (NC) and with coupling systems (Fusion, TARSyn, TeRe). (**b**) Relative mRFP fluorescence of the coupling systems compared to the NC with R^2 values. A perfect correlation indicates a coupling system and reporter which does not influence the expression of the POI.

The coupling systems we tested included a direct fusion using an 18-nucleotide linker to reduce protein misfolding (Chen, Zaro, and Shen, 2013), the TeRe system, and the TARSyn system in *E. coli*. The ideal coupling system preserves the differences in the original expression levels from the 5 selected mRFP sequences, with corresponding expression levels of the associated reporter. If the secondary structures change substantially due to the addition of the reporter protein and the coupling system, the original order of the RFP expression may change. This would be undesired because it means that eventual dissociation of the coupling system after the selection phase would result in shifts in the ranking of POI production. In an industrial setting, where the co-production of a reporter should be avoided, this can result in reduced production after removal of the reporter system.

Upon incorporation of a coupling system and a GFPuv reporter protein in the mRFP-1 to mRFP-5 constructs, an overall downward shift in mRFP production levels is observed (Figure 6.3a). However, the ranking of the mRFP variants does not change, except for the TARSyn system (Figure 6.3b). The TeRe system has the lowest influence on the relative expression ($R^2 = 0.8832$). Generally, all 3

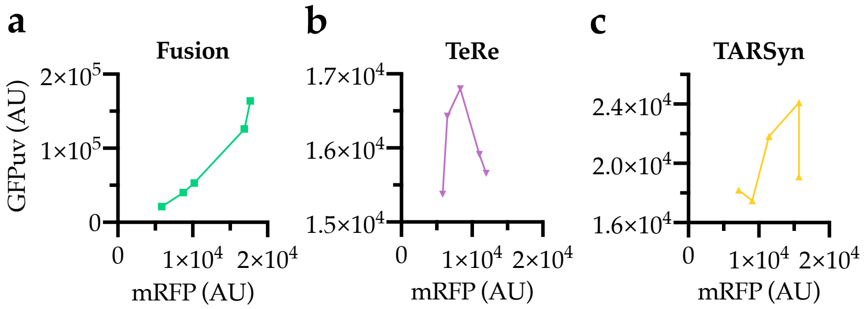


FIGURE 6.4: **Correlation between the reporter (GFPuv) and the POI (mRFP) for different translational coupling systems.** (**a**) A direct fusion of mRFP and GFPuv separated by an 18-nucleotide linker. (**b**) A TeRe system (TGATG). (**c**) The TARSyn system. Note that the overall GFPuv expression for the TeRe and TARSyn systems is about 10-fold lower compared to a direct fusion.

systems have little impact on the relative mRFP expression meaning that the addition (or removal after selection) of the coupling system and reporter does not drastically influence the relative performance. Removal of the coupling system does, however, greatly increase the absolute expression level. This finding that the ranking with/without fusion is similar, agrees with the notion that mainly the translation context in the 5′ part of the CDS influences the overall translation efficiency as shown in Chapter 4. The finding that the overall production of the fused constructs decreases can be explained by changes in mRNA stability.

The correlation between the POI and reporter for a direct fusion is good (Figure 6.4a). A linear relation is expected as long as the mRFP or GFPuv fluorescence is not influenced by the fusion. The mRFP expression is only slightly influenced (Figure 6.3b, R2 0.8486) and there is a good correlation (R^2 0.9595, p = 0.0035) between the mRFP and GFPuv expression. Based on these data it seems that a direct fusion of a reporter to the POI, including a linker, is a good method to indirectly quantify the protein production of the POI. Somewhat surprisingly the addition of a large part of mRNA (the reporter and linker) did not influence the relative expression of the different mRFP codon sequences. Thus, vice versa, if a reporter fusion is used to identify high expressing variants, subsequent removal of the reporter, is not expected to lead to dramatic changes in the relative expression levels of the POI. These results make it likely that a eukaryotic 2A fusion will give similar results as the translation is still coupled 1:1, whereas the proteins are separated by the 2A linker which could be beneficial for certain reporters. A downside of a fusion is that 50% of the production is spent on synthesizing the reporter as the ratio is 1:1. In order to not limit the overall protein production and thereby miss the top producing variants, the transcription should be sufficiently low. As noted earlier, it is generally a good idea to have a low transcription rate when optimizing translation and after optimal translation is established ramp up transcription rates again.

The correlation between the POI and reporter using a TeRe or TARSyn system is largely lost (Figure 6.4b, c). The overall GFPuv expression is about 10-fold lower compared to the fusion system. The GFPuv fluorescence for the TeRe barely rose above the background fluorescence levels of the cells which could explain the poor correlation due to the influence of noise. However, as noted earlier, the correlation of the TeRe system can be highly influenced by the presence of RBS like sequences in the 3' CDS of the POI. Therefore, this system might not be ideal for screening large amounts of codon randomized sequences. The TARSyn system shows a slightly better correlation between the POI and reporter compared to the TeRe and also showed GFPuv fluorescent above the background levels. However, in the case of mRFP-5, the TARSyn coupling was greatly influenced by the codon sequence, and the expression of the reporter suffered. It is possible that the TARSyn secondary structure was disturbed due to a particular nucleotide sequence in the POI. Based on these results it is concluded that the TeRe and TARSyn system is not reliable for the identification of high expressing codon variants.

Reporters such as fluorescent proteins require FACS to identify highly fluorescent variants. An alternative reporter would be an antibiotic resistance marker, coupling high production to cell viability (Figure 6.2c). Antibiotics such as carbenicillin/ampicillin, chloramphenicol, spectinomycin, kanamycin, and tetracycline can be used as a selection marker that can be linked to expression as was demonstrated for the TARSyn system (Rennig et al., 2018). This approach would allow for the selection of a large number of codon variations, and as such is feasible for optimizing protein production processes. A preliminary experiment with a chloramphenicol resistance marker fused to a low and highly expressed protein showed that expression can be coupled to cell viability. The low-expressed mRFP-1 and high-expressed mRFP-5 acted as the extremes of the generated library from Chapter 4. Since the expressed mRFP and fused resistance marker stay in the cytoplasm we needed a resistance marker that also acts in the cytoplasm. We used a chloramphenicol resistance marker (cat) as chloramphenicol inhibits the peptidyl transferase activity of the ribosome and does not cause genomic mutations,

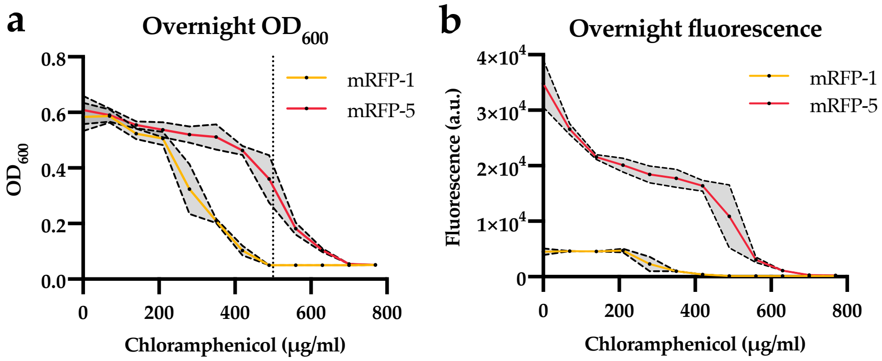


FIGURE 6.5: **Overnight growth results of cells expressing cat-fused mRFP at a low level (mRFP-1) and a high level (mRFP-5).** OD_{600} (**a**) and fluorescence (**b**) measurements of cultures grown overnight. Gray error bands show the standard deviation (n = 4).

thus leaving the selected transcript without mutations. After growing both constructs overnight at different chloramphenicol concentrations, it is clear that the cells with higher mRFP expression (mRFP-5) can tolerate higher concentrations of chloramphenicol (Figure 6.5a). The fluorescence of the cultures confirms that the expression of mRFP-5 is higher (Figure 6.5b, at 0 mg/mL chloramphenicol). Additionally, the effect of the chloramphenicol can be observed as the mRFP fluorescence is reduced at higher concentrations due to an increased rate of protein mutations. At 500 µg/mL chloramphenicol, the high expressing variant would survive unlike the low expressing variant (Figure 6.5a, vertical dotted line). Practically, cells can be transformed with a genetic library and grown without selection pressure to ensure multiple clones of a single construct exist within the population. The cell culture can then be split and plated on a solid medium with an increasing range of antibiotics. Colonies that appear on the plates with a high level of antibiotics are likely to express the POI at high levels. This can be confirmed by removing the antibiotic pressure, measuring the gene expression, and sequencing the responsible sequence.

Screening and selection, as described here, is not limited to codon usage. Promoters and UTRs, or combinations thereof, can also be optimized in this way. Even random or directed modifications on native cellular processes that influence the overall protein production can be indirectly quantified using these systems.

Therefore, randomization, screening and selection offers a solid alternative to *a priori* optimization using predictive tools.

6.2.5 Future prospects for 5′UTR optimization

The predictability and optimization of 5′UTRs in prokaryotes is mainly limited by the reliability of *in silico* predictions of secondary structures. The accessibility and sequence identity of the RBS is the main driving force for the translation initiation rate. Since the two are intrinsically linked, simply using the same RBS for every gene to be expressed will not give constant results. Tools have been developed to predict the translation rates (Bonde et al., 2016; Jeschek, Gerngross, and Panke, 2016; Salis, Mirsky, and Voigt, 2009) but these, in our experience, still have limited reliability when applied to a wide range of proteins. 5′UTRs such as the BCD or fixed 5′ CDS fusions may offer more robust methods to obtain consistent translation rates. The normalizing effect of a constant 5′ CDS (such as a purification tag, which are typically 6-10 amino acids/codons long) matches our findings regarding codon usage in Chapter 4 where we observed that the identity of codon 2 - 9 contributed most to the overall translation efficiency.

6.2.6 Future prospects for 3′UTR optimization

The surprising effect of the sequence between the stop codon and terminator (PSAT region) on translation efficiency in *E. coli* has been highlighted in Chapter 5. This adds an additional tool to the toolbox for the tuning of protein production in this host. However, *E. coli* is an intensively studied microorganism with a wide variety of genetic elements that can be used for tuning translation, so the addition of another tunable element may have limited impact. On the other hand, non-model organisms with less known genetic elements may very well benefit more from this discovery. It will be very interesting to see if the effect of the PSAT region is not only universal for different genes within the same species but also outside the species. If the effect of the PSAT region is due to interactions with more universal prokaryotic elements such as the intrinsic terminator sequence or the ribosomal release mechanisms at the stop codon, there is a possibility that the PSAT regions may also act independent of the expression host, perhaps with the exclusion of organisms that grow at substantially different temperatures such as thermophiles due to temperature-induced changes in secondary structure (Chursov et al., 2013). Testing the tunability and consistency of the PSAT regions in other organisms, including eukaryotes, will help elucidate its fundamental mechanism, and hopefully contribute to a universal tuning mechanism.

Eukaryotic 3'UTRs have already been shown to have great potential for tuning the expression levels. Rational designs of 3'UTRs (including terminator and upstream non-coding region) in combination with non-viral promoters have outperformed the commonly used SV40 genetic elements in HT1080 and HEK293 cell lines (Cheng et al., 2019). Cheng *et al.* showed that an increase in mRNA stability was responsible for the increase in expression. However, they only tested their 3'UTRs for a single reporter gene, so no conclusions can be drawn yet regarding the universality of this region in eukaryotes.

6.2.7 Future prospects for codon optimization

Codon optimization algorithms are continuously being developed and published for many different organisms, and based on many different features. Unfortunately, many algorithms have not been tested *in vivo* in great detail, or the results have not been released. However, only a few algorithms focus on the now generally accepted limitation of prokaryotic translation: secondary structures within the 5'UTR and between the 5'UTR and codons. Algorithms have been developed that optimize the RBS based on predicted interactions with the CDS such as the EMOPEC (Bonde et al., 2016) and RedLibs (Jeschek, Gerngross, and Panke, 2016). While these optimization strategies certainly have had successes, they likely cannot reach the full protein production potential. Ideally, algorithms optimize the 5'UTR and codon sequence simultaneously by using reliable predictions of the mRNA secondary structure. At present, the latter appears a major hurdle.

Another approach to the more rational design is the use of computer learning. In Chapter 4 we have described the potential of these kinds of algorithms. However, in order to generate a robust, general and predictive algorithm, input of a lot of experimental data is required, based on different proteins, and whenever possible in different expression hosts. Only then can these algorithms discover more general rules regarding gene expression. A downside of this approach is that the fundamentals behind gene expression may not directly become clear, but if high expression is the only goal this does not matter. Still, obtained trends most likely will inspire (academic) researchers to eventually uncover the molecular features that govern protein production efficiency.

6.2.8 An efficient optimization strategy

While there is no single optimal strategy for gene optimization for high expression due to protein specific requirements during translation, I suggest the following generalized approach for gene expression in *E.coli* (Figure 6.6). This flow chart shows a step-wise approach with incrementally more involving steps that can be taken to increase the overall protein production. If the wild type sequence is available, I suggest keeping it for the sake of simplicity. Only if the host has a substantially different codon usage compared to the original organism or only the amino acid sequence is available, a codon optimization strategy can be employed. Here, no single best algorithm exists but as long as there is a focus on codons with readily available tRNA's it will be unlikely that translation elongation is the limiting factor. The first step consists of the incorporation of a BCD. This is a relatively easy step and has a high chance for an increase in production (see Chapter 3). If production is not yet sufficient, the next step will be the incorporation of an optimal 3'UTR including a strong synthetic terminator. Due to the CDS-independent influence of this region expression will likely increase, partly due to the unknown effects of the PSAT region and due to the stabilizing effects on the mRNA by the strong terminator. If more expression is required a next relatively easy step is to incorporate a 5' (purification) tag which has been codon optimized in combination with the used BCD. Since the majority of codon usage effects on translation originate from the first 9 codons (see Chapter 4), using an optimized 5' tag will help standardize and improve protein production. An optimized tag that complements a specific BCD can be obtained from a randomization and selection experiment as described previously. These tags can be easily generated prior, using a fluorescent reporter protein and can be reused for different protein production studies. If no tag can be used, e.g. due to protein folding issues, a custom 5'CDS optimization strategy can be employed as described previously, where codons in the 5'CDS are synonymously mutated and an optimal 5'CDS sequence is selected using translational coupling devices. While this overall approach might not be successful for every protein (e.g. toxic proteins or membrane proteins) it offers a simple strategy to increase the success rate for protein production.

6.2.9 A universal high-throughout screening method for protein production

Cell-free protein production is gaining popularity in small-scale expression studies. Cell-to-cell variability is no longer an issue and expression studies can be compared to each other more reliably as long as the same cell-free expression batch is used. Another advantage of cell-free systems is that it allows the synthesis of toxic proteins. A big downside of cell-free expression is the cost and limited scalability. Additionally, not all proteins fold correctly in cell-free systems (e.g. membrane proteins actually need membranes to fold correctly). The natural proliferation capacity of cells makes them ideal, scalable production systems.

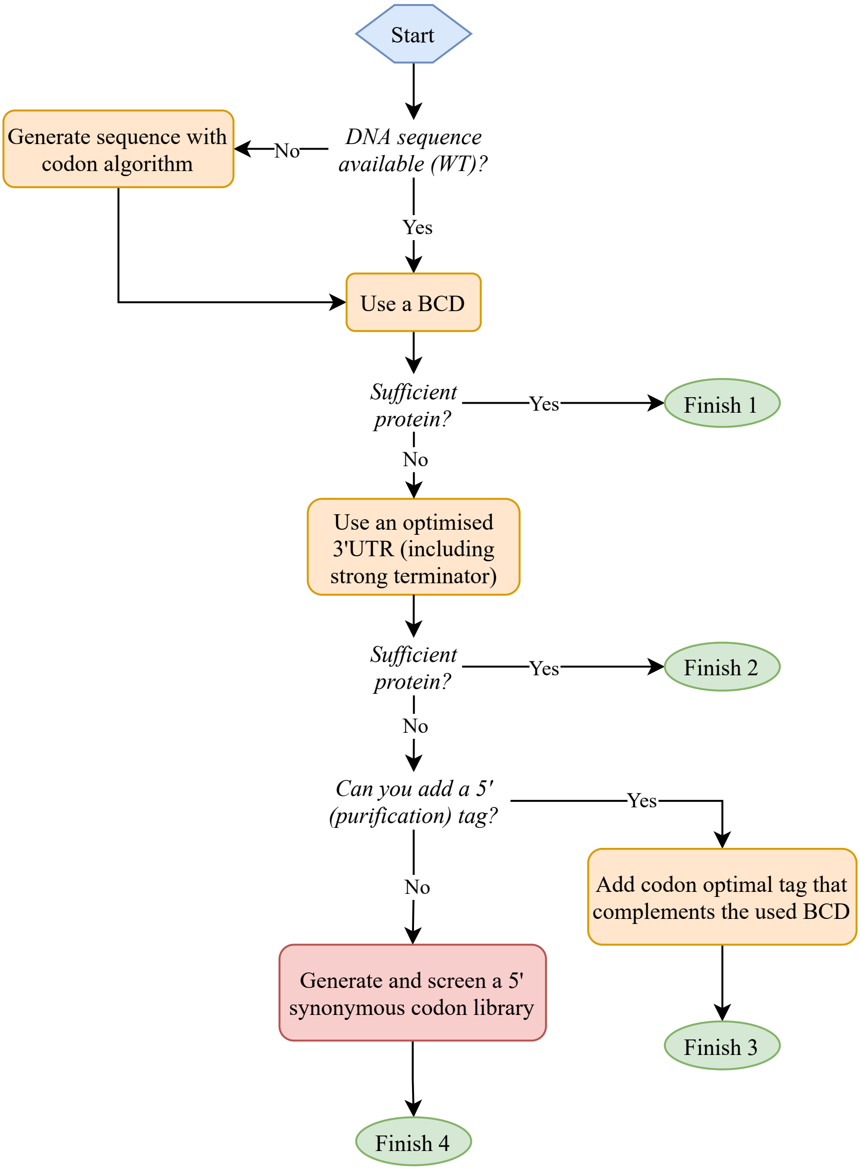

Start
DNA sequence available (WT)?
No
Generate sequence with codon algorithm
Yes
Use a BCD
Sufficient protein?
Yes
Finish 1
No
Use an optimised 3'UTR (including strong terminator)
Sufficient protein?
Yes
Finish 2
No
Can you add a 5′ (purification) tag?
Yes
Add codon optimal tag that complements the used BCD
Finish 3
No
Generate and screen a 5' synonymous codon library
Finish 4

FIGURE 6.6: **A generalized approach to achieve high protein production in *E. coli*.** Steps indicated in orange require relatively low effort. The step in red is much more labour intensive.

Therefore, if large volumes are required (bulk production of enzymes) cell-free lysates are probably not the best approach.

Despite some drawbacks, cell-free systems may offer a great alternative for the screening of large DNA libraries *in vivo* if a selection system would be generated. Degenerate DNA sequences easily reach astronomical numbers of different DNA molecules. However, screening limitations arise due to the limited transformation efficiencies of cells. Cell-free systems have the potential to screen the much higher numbers of the genetic library. However, in order for this to work, two criteria need to be met. First, there should be a decent correlation between expression efficiency *in vitro* compared to *in vivo*. An optimal sequence identified *in vitro* should perform similarly *in vivo* since the large-scale production of proteins *in vivo* is still much more feasible. Second, a way to identify the top-performing sequences needs to be established. The natural link between gene input (genotype; variant sequence) and protein output (phenotype; yield, activity, specificity, stability) that exists in cells (*in vivo*), is lost as all the variant genes and all resulting proteins exist in a single mix in the cellular lysate (*in vitro*). A solution to this is an additional unique peptide tag that can be used to trace back the variant gene that encodes a specific (well performing) protein (Figure 6.7). Quantitative proteomics can be used both to identify the most common and thus most efficiently produced protein, as well as to identify the associated unique peptide tag. Deep DNA sequencing can then be used to obtain the transcript sequences and together with the obtained peptide tag, the responsible transcript sequence can be identified. An approach like this has recently been successfully applied to identify protein variants based on their specific ligand-binding capacity (Egloff et al., 2019). Egloff *et al.* showed that with an 11-15 amino acid tag, they were able to identify the best performing proteins from a pool of 1,000 protein variants using 30,000 unique tag sequences. Their tag excess compared to protein targets was needed to correct for signal variation in the LC-MS/MS measurements. A 30-fold excess does limit the total number of sequences that can be screened. This limitation relates to the sequencing coverage that needs to be reached and the total number of peptides that can be identified in quantitative proteomics. With these limitations, Eglof *et al.* estimate that a total of 13,000-20,000 proteins can be assessed in a single screening experiment. However, if the identification tags would be pre-selected for similar LC-MS/MS signal intensities, the 30-fold excess could be reduced to 1-fold, allowing a single screen to reach up to 600,000 identifiable tags ($30 * 20,000$), in principle allowing quantification of 600,000 different protein variants.

An interesting additional benefit of the screening and selection using LC-MS/MS is that the protein sample can be pre-selected. Meaning that for instance

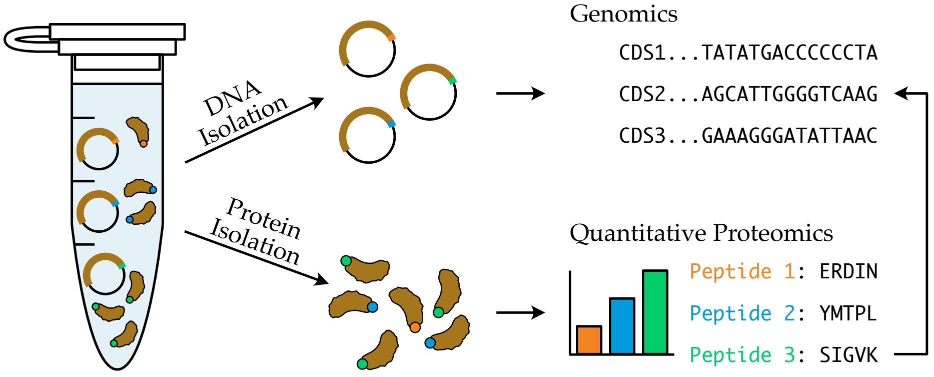


FIGURE 6.7: **Quantitative proteomics and genomics to identify high-producing DNA variants in cell-free systems.** After expression, the DNA and proteins are isolated separately. The protein tags are cleaved from the protein and used in LC-MS/MS quantification. The amino acid sequence of the highest represented tag is used to identify the responsible CDS by matching it to the translated DNA tags from the genomics analysis.

only the soluble fraction can be isolated ensuring that the final selected DNA sequence is both high expressing and leads to a correctly folded protein that is soluble and did not end up in inclusion bodies. Also, stable variants can be selected by incubation at a certain temperature (or in the presence of proteases or denaturing agents); only stable variants will stay in solution. Selection can also be done on protein binding affinity/specificity similarly to the research of Egloff *et al.* (Egloff et al., 2019). Binding assays can be performed for antibodies to ensure that the proteins with their unique tags that will be analyzed in LC-MS/MS are both well expressed and have the intended properties. Additionally, this screening and selection system offers a solution for screening in eukaryotes. Transiently transfected eukaryotic cells can receive multiple different copies of genetic variants and thus the total protein production in that cell cannot be attributed to a specific genetic variant. However, with this tagging approach, the individual protein yield can be examined and can be normalized to the DNA concentrations, which can be obtained via deep sequencing approaches.

For even higher throughput screening, novel techniques can be used to encapsulate the cell lysate mixed with the DNA into microdroplets. Importantly, such an approach circumvents transformation limitations and creates "synthetic cell factories" in the form of droplets (Ma et al., 2021). High expressing DNA molecules can then be screened using a fluorescent reporter linked to the POI and selected by FACS. However, while this approach seems promising it will be difficult to avoid multiple copies being incorporated into a single droplet, giving artificial high protein output related to DNA quantity instead of translation efficiency. However, if the concentrations are tuned right this might be a promising approach to at least enrich for high expressing constructs. Manual screening *in vivo* of the top selected sequences can then help confirm the high translation efficiency of the genetic variant.

6.3 Conclusion

The optimization of mRNA for high protein production remains a complex issue. Advances have been made concerning predictive algorithms, but still, the presently available algorithms can give little guarantee for success. With the rise of computer learning, predictions will improve but possibly at the cost of fundamental understanding. Genetic elements that standardize expression processes, such as the BCD and fixed 5′CDS tags, are gaining popularity as an alternative to rational design. But perhaps more promising is the high throughput screening and selection of optimal sequences from randomized libraries. The gene optimization field is moving quickly by combining different scientific fields and due to the rise of novel techniques. I am personally very excited, and curious to see what comes next!

Appendix A

References

Aksoy, S, C L Squires, and C Squires (1984). "Translational coupling of the trpB and trpA genes in the Escherichia coli tryptophan operon." In: *Journal of Bacteriology* 157.2, pp. 363–367. ISSN: 0021-9193. DOI: 10.1128/JB.157.2.363-367.1984.

Alberts, Bruce et al. (2009). *Essential Cell Biology*. Ed. by Garland Science. third edit. ISBN: 978-0-8153-4130-7.

Andersson, S G and C G Kurland (June 1990). "Codon preferences in free-living microorganisms." In: *Microbiological reviews* 54.2, pp. 198–210. ISSN: 0146-0749.

Angov, Evelina, Patricia M. Legler, and Ryan M. Mease (2011). "Adjustment of Codon Usage Frequencies by Codon Harmonization Improves Protein Expression and Folding". In: *Methods in Molecular Biology*. Vol. 705, pp. 1–13. ISBN: 9781617379666. DOI: 10.1007/978-1-61737-967-3{_}1.

Angov, Evelina et al. (May 2008). "Heterologous Protein Expression Is Enhanced by Harmonizing the Codon Usage Frequencies of the Target Gene with those of the Expression Host". In: *PLoS ONE* 3.5. Ed. by Christophe Herman, e2189. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0002189.

Arango, Daniel et al. (Dec. 2018). "Acetylation of Cytidine in mRNA Promotes Translation Efficiency". In: *Cell* 175.7, pp. 1872–1886. ISSN: 00928674. DOI: 10.1016/j.cell.2018.10.030.

Arribere, Joshua A. et al. (June 2016). "Translation readthrough mitigation". In: *Nature* 534.7609, pp. 719–723. ISSN: 0028-0836. DOI: 10.1038/nature18308.

Baez, William D. et al. (Nov. 2019). "Global analysis of protein synthesis in Flavobacterium johnsoniae reveals the use of Kozak-like sequences in diverse bacteria". In: *Nucleic Acids Research* 47.20, pp. 10477–10488. ISSN: 0305-1048. DOI: 10.1093/nar/gkz855.

Bazzini, Ariel A et al. (Oct. 2016). "Codon identity regulates mRNA stability and translation efficiency during the maternal-to-zygotic transition." In: *The EMBO journal* 35.19, pp. 2087–2103. ISSN: 1460-2075. DOI: 10.15252/embj.201694699.

Bhattacharyya, Sanchari et al. (June 2018). “Accessibility of the Shine-Dalgarno Sequence Dictates N-Terminal Codon Bias in E. coli”. In: *Molecular Cell* 70.5, pp. 894–905. ISSN: 10972765. DOI: 10.1016/j.molcel.2018.05.008.

Boël, Grégory et al. (Jan. 2016). “Codon influence on protein expression in E. coli correlates with mRNA levels”. In: *Nature* 529.7586, pp. 358–363. ISSN: 0028-0836. DOI: 10.1038/nature16509.

Boer, Carl G. de et al. (Jan. 2020). “Deciphering eukaryotic gene-regulatory logic with 100 million random promoters”. In: *Nature Biotechnology* 38.1, pp. 56–65. ISSN: 1087-0156. DOI: 10.1038/s41587-019-0315-8.

Bonde, Mads T. et al. (Mar. 2016). “Predictable tuning of protein expression in bacteria”. In: *Nature Methods* 13.3, pp. 233–236. ISSN: 1548-7091. DOI: 10.1038/nmeth.3727.

Bourret, Jérôme, Samuel Alizon, and Ignacio G. Bravo (Dec. 2019). “COUSIN (COdon Usage Similarity INdex): A Normalized Measure of Codon Usage Preferences”. In: *Genome Biology and Evolution* 11.12. Ed. by Gwenael Piganeau, pp. 3523–3528. ISSN: 1759-6653. DOI: 10.1093/gbe/evz262.

Braun, F. (Aug. 1998). “Ribosomes inhibit an RNase E cleavage which induces the decay of the rpsO mRNA of Escherichiacoli”. In: *The EMBO Journal* 17.16, pp. 4790–4797. ISSN: 14602075. DOI: 10.1093/emboj/17.16.4790.

Brophy, Jennifer AN and Christopher A Voigt (Jan. 2016). “Antisense transcription as a tool to tune gene expression”. In: *Molecular Systems Biology* 12.1, p. 854. ISSN: 1744-4292. DOI: 10.15252/msb.20156540.

Buchan, J. Ross (Feb. 2006). “tRNA properties help shape codon pair preferences in open reading frames”. In: *Nucleic Acids Research* 34.3, pp. 1015–1027. ISSN: 0305-1048. DOI: 10.1093/nar/gkj488.

Buchan, J. Ross and Ian Stansfield (Sept. 2007). “Halting a cellular production line: responses to ribosomal pausing during translation”. In: *Biology of the Cell* 99.9, pp. 475–487. ISSN: 02484900. DOI: 10.1042/BC20070037.

Buhr, Florian et al. (Feb. 2016). “Synonymous Codons Direct Cotranslational Folding toward Different Protein Conformations”. In: *Molecular Cell* 61.3, pp. 341–351. ISSN: 10972765. DOI: 10.1016/j.molcel.2016.01.008.

Burkhardt, David H. et al. (Jan. 2017). “Operon mRNAs are organized into ORF-centric structures that predict translation efficiency”. In: *eLife* 6, e22037. ISSN: 2050-084X. DOI: 10.7554/eLife.22037.

Burow, Dana A. et al. (Aug. 2018). “Attenuated Codon Optimality Contributes to Neural-Specific mRNA Decay in Drosophila”. In: *Cell Reports* 24.7, pp. 1704–1712. ISSN: 22111247. DOI: 10.1016/j.celrep.2018.07.039.

Buschauer, Robert et al. (Apr. 2020). “The Ccr4-Not complex monitors the translating ribosome for codon optimality”. In: *Science* 368.6488, eaay6912. ISSN: 0036-8075. DOI: 10.1126/science.aay6912.

Cambray, Guillaume, Joao C. Guimaraes, and Adam Paul Arkin (Nov. 2018). "Evaluation of 244,000 synthetic sequences reveals design principles to optimize translation in Escherichia coli". In: *Nature Biotechnology* 36.10, pp. 1005–1015. ISSN: 1087-0156. DOI: 10.1038/nbt.4238.

Campbell, Robert E. et al. (June 2002). "A monomeric red fluorescent protein". In: *Proceedings of the National Academy of Sciences* 99.12, pp. 7877–7882. ISSN: 0027-8424. DOI: 10.1073/pnas.082243699.

Carlson, Robert (Dec. 2009). "The changing economics of DNA synthesis". In: *Nature Biotechnology* 27.12, pp. 1091–1094. ISSN: 1087-0156. DOI: 10.1038/nbt1209-1091.

Chaney, Julie L. et al. (May 2017). "Widespread position-specific conservation of synonymous rare codons within coding sequences". In: *PLOS Computational Biology* 13.5. Ed. by Claus O. Wilke, e1005531. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1005531.

Charneski, Catherine A. and Laurence D. Hurst (Mar. 2013). "Positively Charged Residues Are the Major Determinants of Ribosomal Velocity". In: *PLoS Biology* 11.3. Ed. by Harmit S. Malik, e1001508. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1001508.

Chekulaeva, Marina and Markus Landthaler (Sept. 2016). "Eyes on Translation". In: *Molecular Cell* 63.6, pp. 918–925. ISSN: 10972765. DOI: 10.1016/j.molcel.2016.08.031.

Chen, Chunlai et al. (May 2013). "Dynamics of translation by single ribosomes through mRNA secondary structures". In: *Nature Structural & Molecular Biology* 20.5, pp. 582–588. ISSN: 1545-9993. DOI: 10.1038/nsmb.2544.

Chen, Xiaoying, Jennica L. Zaro, and Wei-Chiang Shen (Oct. 2013). "Fusion protein linkers: Property, design and functionality". In: *Advanced Drug Delivery Reviews* 65.10, pp. 1357–1369. ISSN: 0169409X. DOI: 10.1016/j.addr.2012.09.039.

Cheng, Joseph K. et al. (June 2019). "Design and Evaluation of Synthetic Terminators for Regulating Mammalian Cell Transgene Expression". In: *ACS Synthetic Biology* 8.6, pp. 1263–1275. ISSN: 2161-5063. DOI: 10.1021/acssynbio.8b00285.

Choi, Junhong et al. (Feb. 2016). "N6-methyladenosine in mRNA disrupts tRNA selection and translation-elongation dynamics". In: *Nature Structural & Molecular Biology* 23.2, pp. 110–115. ISSN: 1545-9993. DOI: 10.1038/nsmb.3148.

Chou, Hsin-Jung et al. (Dec. 2017). "Transcriptome-wide Analysis of Roles for tRNA Modifications in Translational Regulation". In: *Molecular Cell* 68.5, pp. 978–992. ISSN: 10972765. DOI: 10.1016/j.molcel.2017.11.002.

Chu, Dominique et al. (Jan. 2014). "Translation elongation can control translation initiation on eukaryotic mRNAs". eng. In: *The EMBO Journal* 33.1, pp. 21–34. ISSN: 02614189. DOI: 10.1002/embj.201385651.

Chursov, Andrey et al. (July 2013). “RNAtips: analysis of temperature-induced changes of RNA secondary structure”. In: *Nucleic Acids Research* 41.W1, W486–W491. ISSN: 1362-4962. DOI: 10.1093/nar/gkt486.

Claassens, Nico J. et al. (Sept. 2017). “Improving heterologous membrane protein production in Escherichia coli by combining transcriptional tuning and codon usage algorithms”. In: *PLOS ONE* 12.9. Ed. by Tamir Tuller, e0184355. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0184355.

Claassens, Nico J. et al. (July 2019). “Bicistronic Design-Based Continuous and High-Level Membrane Protein Production in Escherichia coli”. In: *ACS Synthetic Biology* 8.7, pp. 1685–1690. ISSN: 2161-5063. DOI: 10.1021/acssynbio.9b00101.

Costello, Alan et al. (Feb. 2020). “Reinventing the Wheel: Synthetic Circular RNAs for Mammalian Cell Engineering”. In: *Trends in Biotechnology* 38.2, pp. 217–230. ISSN: 01677799. DOI: 10.1016/j.tibtech.2019.07.008.

Crameri, Andreas et al. (1996). “Improved green fluorescent protein by molecular evolution using DNA shuffling”. In: *Nature Biotechnology* 14.3, pp. 315–x1. ISSN: 10870156. DOI: 10.1038/nbt0396-315.

Crick, F (Aug. 1970). “Central dogma of molecular biology”. In: *Nature* 227.5258, pp. 561–3. ISSN: 0028-0836. DOI: 10.1038/227561a0.

Crick, F H (1958). “On protein synthesis”. In: *Symposia of the Society for Experimental Biology* 12, pp. 138–63. ISSN: 0081-1386.

Crick, Francis (1988). *What Mad Persuit: A Personal View of Scientific Discovery*. BasicBooks, p. 182. ISBN: 0-465-09137-5.

Cuperus, Josh T. et al. (Dec. 2017). “Deep learning of the regulatory grammar of yeast 5′ untranslated regions from 500,000 random sequences”. In: *Genome Research* 27.12, pp. 2015–2024. ISSN: 1088-9051. DOI: 10.1101/gr.224964.117.

Curran, Kathleen A. et al. (July 2015). “Short Synthetic Terminators for Improved Heterologous Gene Expression in Yeast”. In: *ACS Synthetic Biology* 4.7, pp. 824–832. ISSN: 2161-5063. DOI: 10.1021/sb5003357.

De Nijs, Yatti, Sofie L. De Maeseneire, and Wim K. Soetaert (Apr. 2020). “5′ untranslated regions: the next regulatory sequence in yeast synthetic biology”. In: *Biological Reviews* 95.2, pp. 517–529. ISSN: 1464-7931. DOI: 10.1111/brv.12575.

Decoene, Thomas et al. (Feb. 2018). “Toward Predictable 5′UTRs in Saccharomyces cerevisiae : Development of a yUTR Calculator”. In: *ACS Synthetic Biology* 7.2, pp. 622–634. ISSN: 2161-5063. DOI: 10.1021/acssynbio.7b00366.

Demain, Arnold L. and Preeti Vaishnav (May 2009). “Production of recombinant proteins by microbes and higher organisms”. In: *Biotechnology Advances* 27.3, pp. 297–306. ISSN: 07349750. DOI: 10.1016/j.biotechadv.2009.01.008.

Deneke, Carlus, Reinhard Lipowsky, and Angelo Valleriani (July 2013). "Effect of ribosome shielding on mRNA stability". In: *Physical Biology* 10.4, p. 046008. ISSN: 1478-3967. DOI: 10.1088/1478-3975/10/4/046008.

Deuschle, U et al. (Nov. 1986). "Promoters of Escherichia coli: a hierarchy of in vivo strength indicates alternate structures." In: *The EMBO journal* 5.11, pp. 2987–94. ISSN: 0261-4189.

Ding, Wentao et al. (Dec. 2018). "Engineering the 5′ UTR-Mediated Regulation of Protein Abundance in Yeast Using Nucleotide Sequence Activity Relationships". In: *ACS Synthetic Biology* 7.12, pp. 2709–2714. ISSN: 2161-5063. DOI: 10.1021/acssynbio.8b00127.

Drummond, D. Allan and Claus O. Wilke (July 2008). "Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution". eng. In: *Cell* 134.2, pp. 341–352. ISSN: 00928674. DOI: 10.1016/j.cell.2008.05.042.

Eden, E. et al. (Feb. 2011). "Proteome Half-Life Dynamics in Living Human Cells". In: *Science* 331.6018, pp. 764–768. ISSN: 0036-8075. DOI: 10.1126/science.1199784.

Edri, Shlomit and Tamir Tuller (July 2014). "Quantifying the Effect of Ribosomal Density on mRNA Stability". In: *PLoS ONE* 9.7. Ed. by Sung Key Jang, e102308. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0102308.

Egloff, Pascal et al. (May 2019). "Engineered peptide barcodes for in-depth analyses of binding protein libraries". In: *Nature Methods* 16.5, pp. 421–428. ISSN: 1548-7091. DOI: 10.1038/s41592-019-0389-8.

Elena, Claudia et al. (2014). "Expression of codon optimized genes in microbial systems: current industrial applications and perspectives". In: *Frontiers in Microbiology* 5.FEB, pp. 1–8. ISSN: 1664-302X. DOI: 10.3389/fmicb.2014.00021.

Espah Borujeni, Amin, Anirudh S. Channarasappa, and Howard M. Salis (Feb. 2014). "Translation rate is controlled by coupled trade-offs between site accessibility, selective RNA unfolding and sliding at upstream standby sites". In: *Nucleic Acids Research* 42.4, pp. 2646–2659. ISSN: 1362-4962. DOI: 10.1093/nar/gkt1139.

Espah Borujeni, Amin and Howard M. Salis (June 2016). "Translation Initiation is Controlled by RNA Folding Kinetics via a Ribosome Drafting Mechanism". In: *Journal of the American Chemical Society* 138.22, pp. 7016–7023. ISSN: 0002-7863. DOI: 10.1021/jacs.6b01453.

Espah Borujeni, Amin et al. (May 2017). "Precise quantification of translation inhibition by mRNA structures that overlap with the ribosomal footprint in N-terminal coding sequences". In: *Nucleic Acids Research* 45.9, pp. 5437–5448. ISSN: 0305-1048. DOI: 10.1093/nar/gkx061.

Fath, Stephan et al. (Mar. 2011). "Multiparameter RNA and Codon Optimization: A Standardized Tool to Assess and Enhance Autologous Mammalian Gene

Expression". In: *PLoS ONE* 6.3. Ed. by Grzegorz Kudla, e17596. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0017596.

Faure, Guilhem et al. (Dec. 2016). "Role of mRNA structure in the control of protein folding". In: *Nucleic Acids Research* 44.22, pp. 10898–10911. ISSN: 0305-1048. DOI: 10.1093/nar/gkw671.

Faure, Guilhem et al. (Dec. 2017). "Adaptation of mRNA structure to control protein folding". In: *RNA Biology* 14.12, pp. 1649–1654. ISSN: 1547-6286. DOI: 10.1080/15476286.2017.1349047.

Ferreira, Joshua P., K. Wesley Overton, and Clifford L. Wang (July 2013). "Tuning gene expression with synthetic upstream open reading frames". In: *Proceedings of the National Academy of Sciences* 110.28, pp. 11284–11289. ISSN: 0027-8424. DOI: 10.1073/pnas.1305590110.

Fleming, Ira and Andre R. O. Cavalcanti (Nov. 2019). "Selection for tandem stop codons in ciliate species with reassigned stop codons". In: *PLOS ONE* 14.11. Ed. by Geoffrey M. Kapler, e0225804. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0225804.

Forrest, Megan E. et al. (Feb. 2020). "Codon and amino acid content are associated with mRNA stability in mammalian cells". In: *PLOS ONE* 15.2. Ed. by Yoon Ki Kim, e0228730. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0228730.

Freitas Nascimento, Janaina de et al. (Mar. 2018). "Codon choice directs constitutive mRNA levels in trypanosomes". In: *eLife* 7, pp. 1–26. ISSN: 2050-084X. DOI: 10.7554/eLife.32467.

Frumkin, Idan et al. (Jan. 2017). "Gene Architectures that Minimize Cost of Gene Expression". In: *Molecular Cell* 65.1, pp. 142–153. ISSN: 10972765. DOI: 10.1016/j.molcel.2016.11.007.

Fu, Jingjing et al. (2016). "Codon usage affects the structure and function of the Drosophila circadian clock protein PERIOD." In: *Genes & development* 30.15, pp. 1761–75. ISSN: 1549-5477. DOI: 10.1101/gad.281030.116.

Fu, Jingjing et al. (Nov. 2018). "Codon usage regulates human KRAS expression at both transcriptional and translational levels". In: *Journal of Biological Chemistry* 293.46, pp. 17929–17940. ISSN: 0021-9258. DOI: 10.1074/jbc.RA118.004908.

Galperin, Michael Y. and Eugene V. Koonin (June 2000). "Who's your neighbor? New computational approaches for functional genomics". In: *Nature Biotechnology* 18.6, pp. 609–613. ISSN: 1087-0156. DOI: 10.1038/76443.

Gamble, Caitlin E. et al. (July 2016). "Adjacent Codons Act in Concert to Modulate Translation Efficiency in Yeast". In: *Cell* 166.3, pp. 679–690. ISSN: 00928674. DOI: 10.1016/j.cell.2016.05.070.

Gardin, Justin et al. (Oct. 2014). "Measurement of average decoding rates of the 61 sense codons in vivo". In: *eLife* 3, pp. 1–20. ISSN: 2050-084X. DOI: 10.7554/eLife.03735.

Geisberg, Joseph V. et al. (Feb. 2014). "Global Analysis of mRNA Isoform Half-Lives Reveals Stabilizing and Destabilizing Elements in Yeast". In: *Cell* 156.4, pp. 812–824. ISSN: 00928674. DOI: 10.1016/j.cell.2013.12.026.

Goeddel, D. V. et al. (Jan. 1979). "Expression in Escherichia coli of chemically synthesized genes for human insulin". In: *Proceedings of the National Academy of Sciences* 76.1, pp. 106–110. ISSN: 0027-8424. DOI: 10.1073/pnas.76.1.106.

Goodman, Daniel B., George M. Church, and Sriram Kosuri (Oct. 2013). "Causes and Effects of N-Terminal Codon Bias in Bacterial Genes". eng. In: *Science* 342.6157, pp. 475–479. ISSN: 0036-8075. DOI: 10.1126/science.1241934.

Gorochowski, Thomas E. et al. (Mar. 2015). "Trade-offs between tRNA abundance and mRNA secondary structure support smoothing of translation elongation rate". In: *Nucleic Acids Research* 43.6, pp. 3022–3032. ISSN: 1362-4962. DOI: 10.1093/nar/gkv199.

Gould, Nathan, Oliver Hendy, and Dimitris Papamichail (Oct. 2014). "Computational Tools and Algorithms for Designing Customized Synthetic Genes". In: *Frontiers in Bioengineering and Biotechnology* 2.OCT, p. 41. ISSN: 2296-4185. DOI: 10.3389/fbioe.2014.00041.

Govantes, F. (Apr. 1998). "Mechanism of translational coupling in the nifLA operon of Klebsiella pneumoniae". In: *The EMBO Journal* 17.8, pp. 2368–2377. ISSN: 14602075. DOI: 10.1093/emboj/17.8.2368.

Griswold, Karl E. et al. (Jan. 2003). "Effects of codon usage versus putative 5′-mRNA structure on the expression of Fusarium solani cutinase in the Escherichia coli cytoplasm". In: *Protein Expression and Purification* 27.1, pp. 134–142. ISSN: 10465928. DOI: 10.1016/S1046-5928(02)00578-8.

Groenke, Nicole et al. (Apr. 2020). "Mechanism of Virus Attenuation by Codon Pair Deoptimization". In: *Cell Reports* 31.4, p. 107586. ISSN: 22111247. DOI: 10.1016/j.celrep.2020.107586.

Grote, Andreas et al. (July 2005). "JCat: a novel tool to adapt codon usage of a target gene to its potential expression host". In: *Nucleic Acids Research* 33.Web Server, W526–W531. ISSN: 0305-1048. DOI: 10.1093/nar/gki376.

Guarneros, G. and C. Portier (Nov. 1990). "Different specificities of ribonuclease II and polynucleotide phosphorylase in 3′mRNA decay". In: *Biochimie* 72.11, pp. 771–777. ISSN: 03009084. DOI: 10.1016/0300-9084(90)90186-K.

Gustafsson, Claes et al. (May 2012). "Engineering genes for predictable protein expression". In: *Protein Expression and Purification* 83.1, pp. 37–46. ISSN: 10465928. DOI: 10.1016/j.pep.2012.02.013.

Gutman, G. A. and G. W. Hatfield (May 1989). "Nonrandom utilization of codon pairs in Escherichia coli." In: *Proceedings of the National Academy of Sciences* 86.10, pp. 3699–3703. ISSN: 0027-8424. DOI: 10.1073/pnas.86.10.3699.

Hanson, Gavin and Jeff Coller (Jan. 2018). “Codon optimality, bias and usage in translation and mRNA decay”. In: *Nature Reviews Molecular Cell Biology* 19.1, pp. 20–30. ISSN: 1471-0072. DOI: 10.1038/nrm.2017.91.

Harigaya, Yuriko and Roy Parker (Dec. 2016). “Analysis of the association between codon optimality and mRNA stability in Schizosaccharomyces pombe”. In: *BMC Genomics* 17.1, p. 895. ISSN: 1471-2164. DOI: 10.1186/s12864-016-3237-6.

Henderson, Kate L. et al. (Apr. 2017). “Mechanism of transcription initiation and promoter escape by E. coli RNA polymerase.” In: *Proceedings of the National Academy of Sciences of the United States of America* 114.15, E3032–E3040. ISSN: 1091-6490. DOI: 10.1073/pnas.1618675114.

Hess, Anne-Katrin et al. (May 2015). “Optimization of Translation Profiles Enhances Protein Expression and Solubility”. In: *PLOS ONE* 10.5. Ed. by Tamir Tuller, e0127039. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0127039.

Hia, Fabian et al. (Nov. 2019). “Codon bias confers stability to human mRNAs.” In: *EMBO reports* 20.11, e48220. ISSN: 1469-3178. DOI: 10.15252/embr.201948220.

Hockenberry, Adam J. et al. (Oct. 2018). “Within-Gene Shine–Dalgarno Sequences Are Not Selected for Function”. In: *Molecular Biology and Evolution* 35.10. Ed. by Deepa Agashe, pp. 2487–2498. ISSN: 0737-4038. DOI: 10.1093/molbev/msy150.

Höllerer, Simon et al. (Dec. 2020). “Large-scale DNA-based phenotypic recording and deep learning enable highly accurate sequence-function mapping”. In: *Nature Communications* 11.1, p. 3551. ISSN: 2041-1723. DOI: 10.1038/s41467-020-17222-4.

Huber, Madeleine et al. (Dec. 2019). “Translational coupling via termination-reinitiation in archaea and bacteria”. In: *Nature Communications* 10.1, p. 4006. ISSN: 2041-1723. DOI: 10.1038/s41467-019-11999-9.

Hughes, Randall A. and Andrew D. Ellington (Jan. 2017). “Synthetic DNA Synthesis and Assembly: Putting the Synthetic in Synthetic Biology”. In: *Cold Spring Harbor Perspectives in Biology* 9.1, a023812. ISSN: 1943-0264. DOI: 10.1101/cshperspect.a023812.

Huter, Paul et al. (Nov. 2017). “Structural Basis for Polyproline-Mediated Ribosome Stalling and Rescue by the Translation Elongation Factor EF-P”. In: *Molecular Cell* 68.3, pp. 515–527. ISSN: 10972765. DOI: 10.1016/j.molcel.2017.10.014.

Huynen, M. (Aug. 2000). “Predicting Protein Function by Genomic Context: Quantitative Evaluation and Qualitative Inferences”. In: *Genome Research* 10.8, pp. 1204–1210. ISSN: 10889051. DOI: 10.1101/gr.10.8.1204.

Ikemura, T. (Jan. 1985). "Codon usage and tRNA content in unicellular and multicellular organisms." In: *Molecular Biology and Evolution* 2.1, pp. 13–34. ISSN: 1537-1719. DOI: 10.1093/oxfordjournals.molbev.a040335.

Ingolia, Nicholas T. et al. (Apr. 2009). "Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling". eng. In: *Science* 324.5924, pp. 218–223. ISSN: 0036-8075. DOI: 10.1126/science.1168978.

Jeacock, Laura, Joana Faria, and David Horn (Mar. 2018). "Codon usage bias controls mRNA and protein abundance in trypanosomatids". In: *eLife* 7, pp. 1–20. ISSN: 2050-084X. DOI: 10.7554/eLife.32496.

Jeschek, Markus, Daniel Gerngross, and Sven Panke (Sept. 2016). "Rationally reduced libraries for combinatorial pathway optimization minimizing experimental effort". In: *Nature Communications* 7.1, p. 11163. ISSN: 2041-1723. DOI: 10.1038/ncomms11163.

Johannsen, Wilhelm (1909). *Elemente der exakten Erblichkeitslehre*. Gustav Fischer Verlag.

Johnson, Grace E. et al. (Sept. 2020). "Functionally uncoupled transcription–translation in Bacillus subtilis". In: *Nature* 585.7823, pp. 124–128. ISSN: 0028-0836. DOI: 10.1038/s41586-020-2638-5.

Jongh, Ronald P.H. de et al. (Feb. 2020). "Designing Eukaryotic Gene Expression Regulation Using Machine Learning". In: *Trends in Biotechnology* 38.2, pp. 191–201. ISSN: 01677799. DOI: 10.1016/j.tibtech.2019.07.007.

Kelsic, Eric D. et al. (Dec. 2016). "RNA Structural Determinants of Optimal Codons Revealed by MAGE-Seq". In: *Cell Systems* 3.6, pp. 563–571. ISSN: 24054712. DOI: 10.1016/j.cels.2016.11.004.

Ki, Mi-Ran and Seung Pil Pack (Mar. 2020). "Fusion tags to enhance heterologous protein expression". In: *Applied Microbiology and Biotechnology* 104.6, pp. 2411–2425. ISSN: 0175-7598. DOI: 10.1007/s00253-020-10402-8.

Kiel, Michael C., Hideko Kaji, and Akira Kaji (Jan. 2007). "Ribosome recycling: An essential process of protein synthesis". In: *Biochemistry and Molecular Biology Education* 35.1, pp. 40–44. ISSN: 14708175. DOI: 10.1002/bmb.6.

Kim, Soo Jung et al. (Apr. 2015). "Translational tuning optimizes nascent protein folding in cells". In: *Science* 348.6233, pp. 444–448. ISSN: 0036-8075. DOI: 10.1126/science.aaa3974.

Kimura, Satoshi, Veerasak Srisuknimit, and Matthew K. Waldor (Oct. 2020). "Probing the diversity and regulation of tRNA modifications". In: *Current Opinion in Microbiology* 57.Figure 2, pp. 41–48. ISSN: 13695274. DOI: 10.1016/j.mib.2020.06.005.

Komarova, Ekaterina S. et al. (July 2020). "Influence of the spacer region between the Shine–Dalgarno box and the start codon for fine-tuning of the translation efficiency in Escherichia coli". In: *Microbial Biotechnology* 13.4, pp. 1254–1261. ISSN: 1751-7915. DOI: 10.1111/1751-7915.13561.

Kozak, Marilyn (1981). “Possible role of flanking nucleotides in recognition of the AUG initiator codon by eukaryotic ribosomes”. In: *Nucleic Acids Research* 9.20, pp. 5233–5252. ISSN: 0305-1048. DOI: 10.1093/nar/9.20.5233.

Kudla, Grzegorz et al. (Apr. 2009). “Coding-Sequence Determinants of Gene Expression in Escherichia coli”. eng. In: *Science* 324.5924, pp. 255–258. ISSN: 0036-8075. DOI: 10.1126/science.1170160.

Kunec, Dusan and Nikolaus Osterrieder (Jan. 2016). “Codon Pair Bias Is a Direct Consequence of Dinucleotide Bias”. In: *Cell Reports* 14.1, pp. 55–67. ISSN: 22111247. DOI: 10.1016/j.celrep.2015.12.011.

Lahtvee, Petri-Jaan et al. (May 2017). “Absolute Quantification of Protein and mRNA Abundances Demonstrate Variability in Gene-Specific Translation Efficiency in Yeast”. In: *Cell Systems* 4.5, pp. 495–504. ISSN: 24054712. DOI: 10.1016/j.cels.2017.03.003.

Lange, Sita J. et al. (July 2012). “Global or local? Predicting secondary structure and accessibility in mRNAs”. In: *Nucleic Acids Research* 40.12, pp. 5215–5226. ISSN: 1362-4962. DOI: 10.1093/nar/gks181.

Lee, Jookyung and Sergei Borukhov (Nov. 2016). “Bacterial RNA Polymerase-DNA Interaction—The Driving Force of Gene Expression and the Target for Drug Action”. In: *Frontiers in Molecular Biosciences* 3.NOV. ISSN: 2296-889X. DOI: 10.3389/fmolb.2016.00073.

Lenstra, Tineke L. et al. (July 2016). “Transcription Dynamics in Living Cells”. In: *Annual Review of Biophysics* 45.1, pp. 25–47. ISSN: 1936-122X. DOI: 10.1146/annurev-biophys-062215-010838.

Leppek, Kathrin, Rhiju Das, and Maria Barna (Mar. 2018). “Functional 5’ UTR mRNA structures in eukaryotic translation regulation and how to find them”. In: *Nature Reviews Molecular Cell Biology* 19.3, pp. 158–174. ISSN: 1471-0072. DOI: 10.1038/nrm.2017.103.

Levin-Karp, Ayelet et al. (June 2013). “Quantifying Translational Coupling in E. coli Synthetic Operons Using RBS Modulation and Fluorescent Reporters”. In: *ACS Synthetic Biology* 2.6, pp. 327–336. ISSN: 2161-5063. DOI: 10.1021/sb400002n.

Levo, Michal et al. (Feb. 2017). “Systematic Investigation of Transcription Factor Activity in the Context of Chromatin Using Massively Parallel Binding and Expression Assays”. In: *Molecular Cell* 65.4, pp. 604–617. ISSN: 10972765. DOI: 10.1016/j.molcel.2017.01.007.

Li, Gene-Wei, Eugene Oh, and Jonathan S. Weissman (Apr. 2012). “The anti-Shine–Dalgarno sequence drives translational pausing and codon choice in bacteria”. eng. In: *Nature* 484.7395, pp. 538–541. ISSN: 0028-0836. DOI: 10.1038/nature10965.

Li, Gene-Wei et al. (Apr. 2014). "Quantifying Absolute Protein Synthesis Rates Reveals Principles Underlying Allocation of Cellular Resources". In: *Cell* 157.3, pp. 624–635. ISSN: 00928674. DOI: 10.1016/j.cell.2014.02.033.

Li, Rui et al. (Apr. 2016). "Effects of cooperation between translating ribosome and RNA polymerase on termination efficiency of the Rho-independent terminator". In: *Nucleic Acids Research* 44.6, pp. 2554–2563. ISSN: 0305-1048. DOI: 10.1093/nar/gkv1285.

Lorenz, Ronny et al. (Dec. 2011). "ViennaRNA Package 2.0". In: *Algorithms for Molecular Biology* 6.1, p. 26. ISSN: 1748-7188. DOI: 10.1186/1748-7188-6-26.

Ma, Jennifer et al. (Dec. 2021). "Microdroplet-based one-step RT-PCR for ultrahigh throughput single-cell multiplex gene expression analysis and rare cell detection". In: *Scientific Reports* 11.1, p. 6777. ISSN: 2045-2322. DOI: 10.1038/s41598-021-86087-4.

Maertens, Barbara et al. (July 2010). "Gene optimization mechanisms: A multigene study reveals a high success rate of full-length human proteins expressed in Escherichia coli". In: *Protein Science* 19.7, pp. 1312–1326. ISSN: 09618368. DOI: 10.1002/pro.408.

Makoff, A.J. and A.E. Smallwood (1990). "The use of two-cistron constructions in improving the expression of a heterologous gene in E.coli". In: *Nucleic Acids Research* 18.7, pp. 1711–1718. ISSN: 0305-1048. DOI: 10.1093/nar/18.7.1711.

McGovern, P. E. et al. (Dec. 2004). "Fermented beverages of pre- and protohistoric China". In: *Proceedings of the National Academy of Sciences* 101.51, pp. 17593–17598. ISSN: 0027-8424. DOI: 10.1073/pnas.0407921102.

Menéndez-Arias, Luis, Alba Sebastián-Martín, and Mar Álvarez (Apr. 2017). "Viral reverse transcriptases". In: *Virus Research* 234, pp. 153–176. ISSN: 01681702. DOI: 10.1016/j.virusres.2016.12.019.

Menendez-Gil, Pilar et al. (Mar. 2020). "Differential evolution in 3'UTRs leads to specific gene expression in Staphylococcus". In: *Nucleic Acids Research* 48.5, pp. 2544–2563. ISSN: 0305-1048. DOI: 10.1093/nar/gkaa047.

Mierendorf, Robert C. et al. (1998). "Expression and Purification of Recombinant Proteins Using the pET System". In: *Molecular Diagnosis of Infectious Diseases*. New Jersey: Humana Press, pp. 257–292. ISBN: 978-1-59259-597-6. DOI: 10.1385/0-89603-485-2:257.

Mignon, Charlotte et al. (May 2018). "Codon harmonization – going beyond the speed limit for protein expression". In: *FEBS Letters* 592.9, pp. 1554–1564. ISSN: 0014-5793. DOI: 10.1002/1873-3468.13046.

Mirzadeh, Kiavash et al. (Sept. 2015). "Enhanced Protein Production in Escherichia coli by Optimization of Cloning Scars at the Vector–Coding Sequence Junction". In: *ACS Synthetic Biology* 4.9, pp. 959–965. ISSN: 2161-5063. DOI: 10.1021/acssynbio.5b00033.

Mishima, Yuichiro and Yukihide Tomari (Mar. 2016). “Codon Usage and 3′ UTR Length Determine Maternal mRNA Stability in Zebrafish”. In: *Molecular Cell* 61.6, pp. 874–885. ISSN: 10972765. DOI: 10.1016/j.molcel.2016.02.027.

Mittal, Pragya et al. (Aug. 2018). “Codon usage influences fitness through RNA toxicity”. In: *Proceedings of the National Academy of Sciences of the United States of America* 115.34, pp. 8639–8644. ISSN: 10916490. DOI: 10.1073/pnas.1810022115.

Mohammad, Fuad et al. (Feb. 2016). “Clarifying the Translational Pausing Landscape in Bacteria by Ribosome Profiling”. In: *Cell Reports* 14.4, pp. 686–694. ISSN: 22111247. DOI: 10.1016/j.celrep.2015.12.073.

Mohanty, Bijoy K. and Sidney R. Kushner (Sept. 2003). “Genomic analysis in Escherichia coli demonstrates differential roles for polynucleotide phosphorylase and RNase II in mRNA abundance and decay”. In: *Molecular Microbiology* 50.2, pp. 645–658. ISSN: 0950382X. DOI: 10.1046/j.1365-2958.2003.03724.x.

Mohanty, Bijoy K. and Sidney R. Kushner (Sept. 2016). “Regulation of mRNA Decay in Bacteria”. In: *Annual Review of Microbiology* 70.1, pp. 25–44. ISSN: 0066-4227. DOI: 10.1146/annurev-micro-091014-104515.

Mordor Intelligence (2020). *PROTEIN EXPRESSION MARKET - GROWTH, TRENDS, COVID-19 IMPACT, AND FORECASTS (2021 - 2026)*. Tech. rep. Mordor Intelligence.

Mordret, Ernest et al. (Aug. 2019). “Systematic Detection of Amino Acid Substitutions in Proteomes Reveals Mechanistic Basis of Ribosome Errors and Selection for Translation Fidelity”. In: *Molecular Cell* 75.3, pp. 427–441. ISSN: 10972765. DOI: 10.1016/j.molcel.2019.06.041.

Morris, David R. and Adam P. Geballe (Dec. 2000). “Upstream Open Reading Frames as Regulators of mRNA Translation”. In: *Molecular and Cellular Biology* 20.23, pp. 8635–8642. ISSN: 1098-5549. DOI: 10.1128/MCB.20.23.8635-8642.2000.

Mossey, Pamela and Anath Das (Jan. 2013). “Expression of Agrobacterium tumefaciens octopine Ti-plasmid virB8 gene is regulated by translational coupling”. In: *Plasmid* 69.1, pp. 72–80. ISSN: 0147619X. DOI: 10.1016/j.plasmid.2012.09.002.

Mugridge, Jeffrey S., Jeff Coller, and John D. Gross (Dec. 2018). “Structural and molecular mechanisms for the control of eukaryotic 5′–3′ mRNA decay”. In: *Nature Structural & Molecular Biology* 25.12, pp. 1077–1085. ISSN: 1545-9993. DOI: 10.1038/s41594-018-0164-z.

Mustoe, Anthony M. et al. (Mar. 2018). “Pervasive Regulatory Functions of mRNA Structure Revealed by High-Resolution SHAPE Probing”. In: *Cell* 173.1, pp. 181–195. ISSN: 00928674. DOI: 10.1016/j.cell.2018.02.034.

Mutalik, Vivek K. et al. (Apr. 2013a). "Precise and reliable gene expression via standard transcription and translation initiation elements". In: *Nature Methods* 10.4, pp. 354–360. ISSN: 1548-7091. DOI: 10.1038/nmeth.2404.

Mutalik, Vivek K. et al. (Apr. 2013b). "Quantitative estimation of activity and quality for collections of functional genetic elements". In: *Nature Methods* 10.4, pp. 347–353. ISSN: 1548-7091. DOI: 10.1038/nmeth.2403.

Narula, Ashrut et al. (Dec. 2019). "Coding regions affect mRNA stability in human cells". In: *RNA* 25.12, pp. 1751–1764. ISSN: 1355-8382. DOI: 10.1261/rna.073239.119.

Navon, Sharon Penias et al. (June 2016). "Amino acid sequence repertoire of the bacterial proteome and the occurrence of untranslatable sequences". In: *Proceedings of the National Academy of Sciences* 113.26, pp. 7166–7170. ISSN: 0027-8424. DOI: 10.1073/pnas.1606518113.

Nedialkova, Danny D. and Sebastian A. Leidel (June 2015). "Optimization of Codon Translation Rates via tRNA Modifications Maintains Proteome Integrity". In: *Cell* 161.7, pp. 1606–1618. ISSN: 00928674. DOI: 10.1016/j.cell.2015.05.022.

Newbury, Sarah F. et al. (Jan. 1987). "Stabilization of translationally active mRNA by prokaryotic REP sequences". In: *Cell* 48.2, pp. 297–310. ISSN: 00928674. DOI: 10.1016/0092-8674(87)90433-8.

Newman, Zachary R et al. (Mar. 2016). "Differences in codon bias and GC content contribute to the balanced expression of TLR7 and TLR9". In: *Proceedings of the National Academy of Sciences* 113.10, E1362–E1371. ISSN: 0027-8424. DOI: 10.1073/pnas.1518976113.

Nicholson, Allen W. (Jan. 2014). "Ribonuclease III mechanisms of double-stranded RNA cleavage". In: *Wiley Interdisciplinary Reviews: RNA* 5.1, pp. 31–48. ISSN: 17577004. DOI: 10.1002/wrna.1195.

Nieuwkoop, Thijs, Nico J. Claassens, and John van der Oost (Jan. 2019). "Improved protein production and codon optimization analyses in Escherichia coli by bicistronic design." In: *Microbial biotechnology* 12.1, pp. 173–179. ISSN: 1751-7915. DOI: 10.1111/1751-7915.13332.

Nieuwkoop, Thijs et al. (Oct. 2020). "The Ongoing Quest to Crack the Genetic Code for Protein Production". In: *Molecular Cell* 80.2, pp. 193–209. ISSN: 10972765. DOI: 10.1016/j.molcel.2020.09.014.

Nirenberg, M. et al. (Jan. 1966). "The RNA Code and Protein Synthesis". In: *Cold Spring Harbor Symposia on Quantitative Biology* 31, pp. 11–24. ISSN: 0091-7451. DOI: 10.1101/SQB.1966.031.01.008.

Nirenberg, M. W. and J. H. Matthaei (Oct. 1961). "The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides". In: *Proceedings of the National Academy of Sciences* 47.10, pp. 1588–1602. ISSN: 0027-8424. DOI: 10.1073/pnas.47.10.1588.

Nørholm, Morten H.H. et al. (Aug. 2013). "Improved production of membrane proteins in Escherichia coli by selective codon substitutions". In: *FEBS Letters* 587.15, pp. 2352–2358. ISSN: 00145793. DOI: 10.1016/j.febslet.2013.05.063.

Oh, Eugene et al. (Dec. 2011). "Selective Ribosome Profiling Reveals the Cotranslational Chaperone Action of Trigger Factor In Vivo". In: *Cell* 147.6, pp. 1295–1308. ISSN: 00928674. DOI: 10.1016/j.cell.2011.10.044.

Oppenheim, D S and C Yanofsky (Aug. 1980). "Translational coupling during expression of the tryptophan operon of Escherichia coli." In: *Genetics* 95.4, pp. 785–95. ISSN: 0016-6731.

O'Reilly, Francis J. et al. (July 2020). "In-cell architecture of an actively transcribing-translating expressome". In: *Science* 369.6503, pp. 554–557. ISSN: 0036-8075. DOI: 10.1126/science.abb3758.

Parret, Annabel HA, Hüseyin Besir, and Rob Meijers (June 2016). "Critical reflections on synthetic gene design for recombinant protein expression". In: *Current Opinion in Structural Biology* 38, pp. 155–162. ISSN: 0959440X. DOI: 10.1016/j.sbi.2016.07.004.

Pechmann, Sebastian and Judith Frydman (Feb. 2013). "Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding". eng. In: *Nature Structural & Molecular Biology* 20.2, pp. 237–243. ISSN: 1545-9993. DOI: 10.1038/nsmb.2466.

Perriman, R and M Ares (Sept. 1998). "Circular mRNA can direct translation of extremely long repeating-sequence proteins in vivo." In: *RNA (New York, N.Y.)* 4.9, pp. 1047–54. ISSN: 1355-8382. DOI: 10.1017/s135583829898061x.

Peterman, Neil and Erel Levine (Dec. 2016). "Sort-seq under the hood: implications of design choices on large-scale characterization of sequence-function relations". In: *BMC Genomics* 17.1, p. 206. ISSN: 1471-2164. DOI: 10.1186/s12864-016-2533-5.

Petersen, Søren D. et al. (Aug. 2018). "Modular 5′-UTR hexamers for context-independent tuning of protein expression in eukaryotes". In: *Nucleic Acids Research* 46.21, e127. ISSN: 0305-1048. DOI: 10.1093/nar/gky734.

Portin, Petter and Adam Wilkins (Apr. 2017). "The Evolving Definition of the Term "Gene"". In: *Genetics* 205.4, pp. 1353–1364. ISSN: 1943-2631. DOI: 10.1534/genetics.116.196956.

Potapov, Vladimir et al. (Nov. 2018). "Comprehensive Profiling of Four Base Overhang Ligation Fidelity by T4 DNA Ligase and Application to DNA Assembly". In: *ACS Synthetic Biology* 7.11, pp. 2665–2674. ISSN: 2161-5063. DOI: 10.1021/acssynbio.8b00333.

Presnyak, Vladimir et al. (Mar. 2015). "Codon optimality is a major determinant of mRNA stability". In: *Cell* 160.6, pp. 1111–1124. ISSN: 10974172. DOI: 10.1016/j.cell.2015.02.029.

Puigbo, P. et al. (May 2007). "OPTIMIZER: a web server for optimizing the codon usage of DNA sequences". In: *Nucleic Acids Research* 35.Web Server, W126–W131. ISSN: 0305-1048. DOI: 10.1093/nar/gkm219.

Quax, Tessa E.F. et al. (Sept. 2013). "Differential Translation Tunes Uneven Production of Operon-Encoded Proteins". In: *Cell Reports* 4.5, pp. 938–944. ISSN: 22111247. DOI: 10.1016/j.celrep.2013.07.049.

Quax, Tessa E.F. et al. (July 2015). "Codon Bias as a Means to Fine-Tune Gene Expression". In: *Molecular Cell* 59.2, pp. 149–161. ISSN: 10972765. DOI: 10.1016/j.molcel.2015.05.035.

Raab, David et al. (Sept. 2010). "The GeneOptimizer Algorithm: using a sliding window approach to cope with the vast sequence space in multiparameter DNA sequence optimization". In: *Systems and Synthetic Biology* 4.3, pp. 215–225. ISSN: 1872-5325. DOI: 10.1007/s11693-010-9062-3.

Radhakrishnan, Aditya et al. (Sept. 2016). "The DEAD-Box Protein Dhh1p Couples mRNA Decay and Translation by Monitoring Codon Optimality". In: *Cell* 167.1, pp. 122–132. ISSN: 00928674. DOI: 10.1016/j.cell.2016.08.053.

Ranaghan, Matthew J. et al. (Dec. 2021). "Assessing optimal: inequalities in codon optimization algorithms". In: *BMC Biology* 19.1, p. 36. ISSN: 1741-7007. DOI: 10.1186/s12915-021-00968-8.

Reis, Mario dos, Lorenz Wernisch, and Renos Savva (Dec. 2003). "Unexpected correlations between gene expression and codon usage bias from microarray data for the whole Escherichia coli K-12 genome". In: *Nucleic Acids Research* 31.23, pp. 6976–6985. ISSN: 03051048. DOI: 10.1093/nar/gkg897.

Rennig, Maja et al. (Feb. 2018). "TARSyn: Tunable Antibiotic Resistance Devices Enabling Bacterial Synthetic Evolution and Protein Production". In: *ACS Synthetic Biology* 7.2, pp. 432–442. ISSN: 2161-5063. DOI: 10.1021/acssynbio.7b00200.

Rex, G et al. (July 1994). "The mechanism of translational coupling in Escherichia coli. Higher order structure in the atpHA mRNA acts as a conformational switch regulating the access of de novo initiating ribosomes." In: *The Journal of biological chemistry* 269.27, pp. 18118–27. ISSN: 0021-9258.

Rosano, GermÃ¡n L. and Eduardo A. Ceccarelli (Apr. 2014). "Recombinant protein expression in Escherichia coli: advances and challenges". In: *Frontiers in Microbiology* 5.APR, pp. 1–17. ISSN: 1664-302X. DOI: 10.3389/fmicb.2014.00172.

Rouskin, Silvi et al. (Jan. 2014). "Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo". In: *Nature* 505.7485, pp. 701–705. ISSN: 0028-0836. DOI: 10.1038/nature12894.

Roy, Varnika et al. (Aug. 2017). "A bicistronic vector with destabilized mRNA secondary structure yields scalable higher titer expression of human neurturin in E. coli". In: *Biotechnology and Bioengineering* 114.8, pp. 1753–1761. ISSN: 00063592. DOI: 10.1002/bit.26299.

Ryan, M. D., A. M. Q. King, and G. P. Thomas (Nov. 1991). "Cleavage of foot-and-mouth disease virus polyprotein is mediated by residues located within a 19 amino acid sequence". In: *Journal of General Virology* 72.11, pp. 2727–2732. ISSN: 0022-1317. DOI: 10.1099/0022-1317-72-11-2727.

Ryan, M D et al. (Jan. 1997). "The cleavage activities of aphthovirus and cardiovirus 2A proteins." In: *Journal of General Virology* 78.1, pp. 13–21. ISSN: 0022-1317. DOI: 10.1099/0022-1317-78-1-13.

Saito, Kazuki, Rachel Green, and Allen R. Buskirk (Feb. 2020). "Translational initiation in E. coli occurs at the correct sites genome-wide in the absence of mRNA-rRNA base-pairing". In: *eLife* 9, pp. 1–19. ISSN: 2050-084X. DOI: 10.7554/eLife.55002.

Salis, Howard M., Ethan A. Mirsky, and Christopher A. Voigt (Oct. 2009). "Automated design of synthetic ribosome binding sites to control protein expression". In: *Nature Biotechnology* 27.10, pp. 946–950. ISSN: 1087-0156. DOI: 10.1038/nbt.1568.

Salzberg, Steven L. (Dec. 2018). "Open questions: How many genes do we have?" In: *BMC Biology* 16.1, p. 94. ISSN: 1741-7007. DOI: 10.1186/s12915-018-0564-x.

Schmid, Manfred and Torben Heick Jensen (Aug. 2018). "Controlling nuclear RNA levels". In: *Nature Reviews Genetics* 19.8, pp. 518–529. ISSN: 1471-0056. DOI: 10.1038/s41576-018-0013-2.

Schümperli, Daniel et al. (Oct. 1982). "Translational coupling at an intercistronic boundary of the Escherichia coli galactose operon". In: *Cell* 30.3, pp. 865–871. ISSN: 00928674. DOI: 10.1016/0092-8674(82)90291-4.

Shah, Premal et al. (June 2013). "Rate-Limiting Steps in Yeast Protein Translation". In: *Cell* 153.7, pp. 1589–1601. ISSN: 00928674. DOI: 10.1016/j.cell.2013.05.049.

Sharp, Paul M. and Wen-Hsiung Li (1987). "The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications". In: *Nucleic Acids Research* 15.3, pp. 1281–1295. ISSN: 0305-1048. DOI: 10.1093/nar/15.3.1281.

Shine, J. and L. Dalgarno (Apr. 1974). "The 3'-Terminal Sequence of Escherichia coli 16S Ribosomal RNA: Complementarity to Nonsense Triplets and Ribosome Binding Sites". In: *Proceedings of the National Academy of Sciences* 71.4, pp. 1342–1346. ISSN: 0027-8424. DOI: 10.1073/pnas.71.4.1342.

Siegfried, Nathan A. et al. (Sept. 2014). "RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP)". In: *Nature Methods* 11.9, pp. 959–965. ISSN: 1548-7091. DOI: 10.1038/nmeth.3029.

Smit, M. H. de and J. van Duin (Oct. 1990). "Secondary structure of the ribosome binding site determines translational efficiency: a quantitative analysis." In: *Proceedings of the National Academy of Sciences* 87.19, pp. 7668–7672. ISSN: 0027-8424. DOI: 10.1073/pnas.87.19.7668.

Söll, D. et al. (Aug. 1966). "Specificity of sRNA for recognition of codons as studied by the ribosomal binding technique". In: *Journal of Molecular Biology* 19.2, pp. 556–573. ISSN: 00222836. DOI: 10.1016/S0022-2836(66)80023-2.

Studer, Sean M. and Simpson Joseph (Apr. 2006). "Unfolding of mRNA Secondary Structure by the Bacterial Translation Initiation Complex". In: *Molecular Cell* 22.1, pp. 105–115. ISSN: 10972765. DOI: 10.1016/j.molcel.2006.02.014.

Sumant, Onkar and Sohail Shaikh (2017). *Protein Therapeutics Market*. Tech. rep. Allied Market Research, p. 211.

Sun, Manman et al. (Dec. 2020). "Enhanced production of recombinant proteins in Corynebacterium glutamicum by constructing a bicistronic gene expression system". In: *Microbial Cell Factories* 19.1, p. 113. ISSN: 1475-2859. DOI: 10.1186/s12934-020-01370-9.

Takyar, Seyedtaghi, Robyn P. Hickerson, and Harry F. Noller (Jan. 2005). "mRNA Helicase Activity of the Ribosome". In: *Cell* 120.1, pp. 49–58. ISSN: 00928674. DOI: 10.1016/j.cell.2004.11.042.

Taniguchi, Yuichi et al. (July 2010). "Quantifying E. coli Proteome and Transcriptome with Single-Molecule Sensitivity in Single Cells". In: *Science* 329.5991, pp. 533–538. ISSN: 0036-8075. DOI: 10.1126/science.1188308.

Tate, Warren P., Andrew G. Cridge, and Chris M. Brown (Dec. 2018). "'Stop' in protein synthesis is modulated with exquisite subtlety by an extended RNA translation signal". In: *Biochemical Society Transactions* 46.6, pp. 1615–1625. ISSN: 0300-5127. DOI: 10.1042/BST20180190.

Tesina, Petr et al. (Feb. 2020). "Molecular mechanism of translational stalling by inhibitory codon combinations and poly(A) tracts". In: *The EMBO Journal* 39.3, pp. 1–17. ISSN: 0261-4189. DOI: 10.15252/embj.2019103365.

Thompson, Mary K. and Wendy V. Gilbert (Aug. 2017). "mRNA length-sensing in eukaryotic translation: reconsidering the "closed loop" and its implications for translational control". In: *Current Genetics* 63.4, pp. 613–620. ISSN: 0172-8083. DOI: 10.1007/s00294-016-0674-3.

Tian, Bin and James L. Manley (Jan. 2017). "Alternative polyadenylation of mRNA precursors". In: *Nature Reviews Molecular Cell Biology* 18.1, pp. 18–30. ISSN: 1471-0072. DOI: 10.1038/nrm.2016.116.

Torrent, Marc et al. (Sept. 2018). “Cells alter their tRNA abundance to selectively regulate protein synthesis during stress conditions”. In: *Science Signaling* 11.546, eaat6409. ISSN: 1945-0877. DOI: 10.1126/scisignal.aat6409.

Traverse, Charles C. and Howard Ochman (Mar. 2016). “Conserved rates and patterns of transcription errors across bacterial growth states and lifestyles”. In: *Proceedings of the National Academy of Sciences* 113.12, pp. 3311–3316. ISSN: 0027-8424. DOI: 10.1073/pnas.1525329113.

Tuller, Tamir and Hadas Zur (Jan. 2015). “Multiple roles of the coding sequence 5′ end in gene expression regulation”. In: *Nucleic Acids Research* 43.1, pp. 13–28. ISSN: 1362-4962. DOI: 10.1093/nar/gku1313.

Tuller, Tamir et al. (Apr. 2010). “An Evolutionarily Conserved Mechanism for Controlling the Efficiency of Protein Translation”. In: *Cell* 141.2, pp. 344–354. ISSN: 00928674. DOI: 10.1016/j.cell.2010.03.031.

Urtecho, Guillaume et al. (Mar. 2019). “Systematic Dissection of Sequence Elements Controlling σ70 Promoters Using a Genomically Encoded Multiplexed Reporter Assay in Escherichia coli”. In: *Biochemistry* 58.11, pp. 1539–1551. ISSN: 0006-2960. DOI: 10.1021/acs.biochem.7b01069.

Urtecho, Guillaume et al. (2020). “Genome-wide Functional Characterization of Escherichia coli Promoters and Regulatory Elements Responsible for their Function”. In: *bioRxiv*, p. 2020.01.04.894907. DOI: 10.1101/2020.01.04.894907.

Vasquez, John R. et al. (Mar. 1989). “An expression system for trypsin”. In: *Journal of Cellular Biochemistry* 39.3, pp. 265–276. ISSN: 0730-2312. DOI: 10.1002/jcb.240390306.

Vasquez, Kevin A. et al. (Feb. 2016). “Slowing Translation between Protein Domains by Increasing Affinity between mRNAs and the Ribosomal Anti-Shine–Dalgarno Sequence Improves Solubility”. In: *ACS Synthetic Biology* 5.2, pp. 133–145. ISSN: 2161-5063. DOI: 10.1021/acssynbio.5b00193.

Verma, Manasvi et al. (Dec. 2019). “A short translational ramp determines the efficiency of protein synthesis”. In: *Nature Communications* 10.1, p. 5774. ISSN: 2041-1723. DOI: 10.1038/s41467-019-13810-1.

Vytvytska, O et al. (May 2000). “Hfq (HF1) stimulates ompA mRNA decay by interfering with ribosome binding.” In: *Genes & development* 14.9, pp. 1109–18. ISSN: 0890-9369.

Wade, Joseph T. and Kevin Struhl (Apr. 2008). “The transition from transcriptional initiation to elongation”. In: *Current Opinion in Genetics and Development* 18.2, pp. 130–136. ISSN: 0959437X. DOI: 10.1016/j.gde.2007.12.008.

Wan, Ji et al. (Sept. 2018). “A Coding Sequence-Embedded Principle Governs Translational Reading Frame Fidelity”. In: *Research* 2018, pp. 1–15. ISSN: 2639-5274. DOI: 10.1155/2018/7089174.

Watson, James D. (1965). *Molecular Biology of the Gene*. New York: W.A. Benjamin, Inc.

Webster, Michael W. et al. (June 2018). "mRNA Deadenylation Is Coupled to Translation Rates by the Differential Activities of Ccr4-Not Nucleases". In: *Molecular Cell* 70.6, pp. 1089–1100. ISSN: 10972765. DOI: 10.1016/j.molcel.2018.05.033.

Weenink, Tim et al. (Jan. 2018). "Design of RNA hairpin modules that predictably tune translation in yeast". In: *Synthetic Biology* 3.1, pp. 1–9. ISSN: 2397-7000. DOI: 10.1093/synbio/ysy019.

Weinberg, David E. et al. (Feb. 2016). "Improved Ribosome-Footprint and mRNA Measurements Provide Insights into Dynamics and Regulation of Yeast Translation". In: *Cell Reports* 14.7, pp. 1787–1799. ISSN: 22111247. DOI: 10.1016/j.celrep.2016.01.043.

Welch, Mark et al. (Sept. 2009). "Design Parameters to Control Synthetic Gene Expression in Escherichia coli". In: *PLoS ONE* 4.9. Ed. by Grzegorz Kudla, e7002. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0007002.

Wendt, Diane (2013). *Two tons of pig parts: Making insulin in the 1920s*.

Wesselhoeft, R. Alexander, Piotr S. Kowalski, and Daniel G. Anderson (Dec. 2018). "Engineering circular RNA for potent and stable translation in eukaryotic cells". In: *Nature Communications* 9.1, p. 2629. ISSN: 2041-1723. DOI: 10.1038/s41467-018-05096-6.

West, Steven and Nicholas J. Proudfoot (Feb. 2016). "Transcriptional Termination Enhances Protein Expression in Human Cells". In: *Molecular Cell* 61.3, p. 486. ISSN: 10972765. DOI: 10.1016/j.molcel.2016.01.021.

Wilson, Daniel N., Stefan Arenz, and Roland Beckmann (Apr. 2016). "Translation regulation via nascent polypeptide-mediated ribosome stalling". In: *Current Opinion in Structural Biology* 37, pp. 123–133. ISSN: 0959440X. DOI: 10.1016/j.sbi.2016.01.008.

Winkelman, Jared T. et al. (Mar. 2016). "Multiplexed protein-DNA cross-linking: Scrunching in transcription start site selection". In: *Science* 351.6277, pp. 1090–1093. ISSN: 0036-8075. DOI: 10.1126/science.aad6881.

Wu, Qiushuang et al. (Apr. 2019). "Translation affects mRNA stability in a codon-dependent manner in human cells". In: *eLife* 8, pp. 1–22. ISSN: 2050-084X. DOI: 10.7554/eLife.45396.

Yan, Xiaowei et al. (May 2016). "Dynamics of Translation of Single mRNA Molecules In Vivo". In: *Cell* 165.4, pp. 976–989. ISSN: 00928674. DOI: 10.1016/j.cell.2016.04.034.

Yang, Qian et al. (Sept. 2019). "eRF1 mediates codon usage effects on mRNA translation efficiency through premature termination at rare codons". In: *Nucleic Acids Research* 47.17, pp. 9243–9258. ISSN: 0305-1048. DOI: 10.1093/nar/gkz710.

Yona, Avihu H., Eric J. Alm, and Jeff Gore (Dec. 2018). "Random sequences rapidly evolve into de novo promoters". In: *Nature Communications* 9.1, p. 1530. ISSN: 2041-1723. DOI: 10.1038/s41467-018-04026-w.

Yu, Chien-Hung et al. (Sept. 2015). "Codon Usage Influences the Local Rate of Translation Elongation to Regulate Co-translational Protein Folding". In: *Molecular Cell* 59.5, pp. 744–754. ISSN: 10972765. DOI: 10.1016/j.molcel.2015.07.018.

Zhang, Gong, Magdalena Hubalewska, and Zoya Ignatova (Mar. 2009). "Transient ribosomal attenuation coordinates protein synthesis and co-translational folding". In: *Nature Structural & Molecular Biology* 16.3, pp. 274–280. ISSN: 1545-9993. DOI: 10.1038/nsmb.1554.

Zhao, Boxuan Simen, Ian A. Roundtree, and Chuan He (Jan. 2017). "Post-transcriptional gene regulation by mRNA modifications". In: *Nature Reviews Molecular Cell Biology* 18.1, pp. 31–42. ISSN: 1471-0072. DOI: 10.1038/nrm.2016.132.

Zhao, Fangzhou, Chien-hung Yu, and Yi Liu (Aug. 2017). "Codon usage regulates protein structure and function by affecting translation elongation speed in Drosophila cells". In: *Nucleic Acids Research* 45.14, pp. 8484–8492. ISSN: 0305-1048. DOI: 10.1093/nar/gkx501.

Zhao, Ju Ping et al. (Dec. 2018). "AU-rich long 3' untranslated region regulates gene expression in bacteria". In: *Frontiers in Microbiology* 9.DEC, pp. 1–10. ISSN: 1664302X. DOI: 10.3389/fmicb.2018.03080.

Zhou, Kang et al. (2011). "Novel reference genes for quantifying transcriptional responses of Escherichia coli to protein overexpression by quantitative PCR". In: *BMC Molecular Biology* 12.1, p. 18. ISSN: 1471-2199. DOI: 10.1186/1471-2199-12-18.

Zhou, Mian et al. (Mar. 2013). "Non-optimal codon usage affects expression, structure and function of clock protein FRQ". eng. In: *Nature* 495.7439, pp. 111–115. ISSN: 0028-0836. DOI: 10.1038/nature11833.

Zhou, Mian et al. (Sept. 2015). "Nonoptimal codon usage influences protein structure in intrinsically disordered regions". In: *Molecular Microbiology* 97.5, pp. 974–987. ISSN: 0950382X. DOI: 10.1111/mmi.13079.

Zhou, Zhipeng et al. (Oct. 2016). "Codon usage is an important determinant of gene expression levels largely through its effects on transcription". In: *Proceedings of the National Academy of Sciences* 113.41, E6117–E6125. ISSN: 0027-8424. DOI: 10.1073/pnas.1606724113.

Zuker, M. (July 2003). "Mfold web server for nucleic acid folding and hybridization prediction". In: *Nucleic Acids Research* 31.13, pp. 3406–3415. ISSN: 1362-4962. DOI: 10.1093/nar/gkg595.

Zulkower, Valentin and Susan Rosser (Aug. 2020). "DNA Chisel, a versatile sequence optimizer". In: *Bioinformatics* 36.16. Ed. by Yann Ponty, pp. 4508–4509. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btaa558.

Appendix B

Acknowledgements

With this thesis, my PhD is officially ended. Doing a PhD was the most enjoyable part of my education. It truly transformed me from a student to an independent researcher. Of course, there are many people to thank as they all contributed, in one way or another, to this thesis.

John, heel erg bedankt voor alles! Je bent voor mij een geweldige begeleider geweest. Ik ben zeer dankbaar voor je unieke vertrouwen die je hebt in je PhD's waardoor ik echt het gevoel gehad heb zelf m'n eigen onderzoek te doen. Natuurlijk stond je ook altijd klaar met advies mocht dit nodig zijn. Vanaf m'n master thesis vond ik de discussies over codon gebruik met jou al geweldig. Ik denk dat we met z'n allen een mooie toevoeging hebben gedaan aan het begrijpen van codon bias, ook al zal dit fenomeen misschien nooit helemaal begrepen worden. Ik ben verder ook heel erg dankbaar voor het vervolg project wat we samen gaan doen en kijk daar erg naar uit!

Nico, ook al was je officieel pas veel later m'n tweede begeleider, wil ik je ook nog bedanken voor je advies en hulp in m'n 1e jaar toen je nog een mede PhD student was. Het was altijd fijn om met jou te overleggen aangezien je zelf ook kennis en ervaring hebt opgedaan op het gebied van codons. Toen je later ook nog eens m'n officiële tweede begeleider werd was het helemaal mooi. Ik vind je een hele goede begeleider en je lijkt je rol als assistant professor snel op te pakken. Ik hoop dat we samen ons onderzoek kunnen voortzetten in de toekomst en heel veel succes met je verdere carrière.

Joep and **Max**, thank you so much for being my paranymphs, colleagues and friends. We were the original BaSyC boys, and now we can look back on the synthetic cell that we've made together, or can we...? In any case, I especially look back with fondness to the BaSyC related activities we've done together, like sharing a sauna with beers, which is, of course, essential for good science. Thanks again, and let's stay in touch!

Belen, **Janneke**, **Joep**, **Jurre**, **Lorrenzo** and **Max**, you were really great office mates! I am glad we could work hard and have so much fun in our office. The dinners and "The Office"/movie nights we did were great and I hope we can still do some more. One day I will bring that cake since I was so bad at playing "the floor is lava"...

Barbara, je was echt een geweldige en unieke student om te begeleiden. Ik had nog nooit met iemand samen gewerkt die zo snel zelfstandig aan de slag kon en zo snel nieuwe informatie op nam. Ik ben blij dat je na je master thesis een PhD bent gaan doen bij Bioinformatics en dat we daardoor nog een zeer mooi project samen hebben kunnen doen, bedankt daarvoor. Ik wens je het allerbeste met je verdere carrière!

Costas, thank you very much for your enthousiasm towards my project. Interactions with you kindled a desire in me to try to do more with the technology we are developing. I hope we can continue together with this research and whatever is going to happen after, it will be great!

Sjoerd, heel erg bedankt voor al je kennis en hulp over de jaren. Zowel tijdens mijn master thesis als tijdens m'n PhD. Ik kijk uit naar het paper wat we nog samen gaan publiceren.

Raymond, ik zat nog even te twijfelen of ik de moeite moest nemen om voor je een persoonlijke alinea (wat je als eis stelde) te formuleren die uit 1 zin bestaat, maar ik wilde toch ook echt iets meer zeggen. Ons contact was niet zo zeer over biotechnologie maar verrassend over computer wetenschappen en "speed runs". Ik vond het altijd heel leuk om het met je daar over te hebben en ben blij dat we elkaar wat goede Youtube kanalen hebben kunnen adviseren.

Of course I want to thank everyone at BacGen (**Adiini**, **Anneleen**, **Belen**, **Carina**, **Catarina**, **Costas**, **Despoina**, **Eric**, **Eugenios**, **Guus**, **Hanne**, **Ioannis**, **Isabelle**, **Ismael**, **James**, **Janneke**, **Jeroen**, **Joep**, **Jorrit**, **Joyshree**, **Jurre**, **Lorenzo**, **Maartje**, **Mamou**, **Max**, **Miguel**, **Mihris**, **Prarthana**, **Ricardo**, **Rob**, **Serve**, **Stijn**, **Suzan**, **Thomas**, **Wen** and everyone I am forgetting) and everyone at Microbiology for the nice company and amazing working environment. I wish you all the best.

Tom, **Rob**, **Philippe** en **Guus**, heel erg bedankt voor al jullie ondersteunend werk!

Natuurlijk wil ik **Thijs E.**, **Willem**, **Anja**, **Heidi** en **Hannie** bedanken voor al hun werk achter de schermen. Wat jullie doen maakt het uiteindelijk mogelijk dat wij PhD studenten ons kunnen ontwikkelen en mooie wetenschappelijke resultaten kunnen genereren.

Alba, **Anthonie**, **Barbara**, **Sophie**, **Stéphanie**, **Thijmen** and **Yvette**, you were all great students and I really want to thank you all for the work you put in to create a nice thesis for yourselves and help me with my overall project.

Carina, **Jolanda**, **Marie-Luise**, **Max**, **Nicolas**, **Peter** and **Timon**, I think we made some nice changes as a PhD-board towards the general atmosphere and well-being of PhD students within Microbiology. Thank you all for this.

Caifang, **Costas**, **Enrique**, **Ioannis**, **Ivette**, **Lot**, **Ran** and **Thanaporn**, thank you for organizing the PhD trip to the United States. It was an unforgettable trip!

Bob, **Patrick**, **Kasra**, **Stephanie**, **Kevin** en **Ralph**, bedankt voor jullie interesse, discussies en steun. Jullie zijn allemaal geweldige vrienden!

Pap en **Mam**, heel erg bedankt voor jullie steun en constante interesse in wat ik doe. Het is niet altijd even makkelijk voor mij om uit te leggen waar ik precies aan werk, maar jullie delen dezelfde fascinatie en verbijstering met mij als ik uit leg wat een cel allemaal kan doen en dat maakt het toch steeds super leuk om over m'n werk te praten. **Daan** en **Jasmijn**, voor jullie geldt natuurlijk hetzelfde. Ik ben blij dat jullie nu eindelijk een keer goed kunnen zien wat ik precies allemaal heb zitten uitvoeren op die universiteit. **Leo**, **Beppie**, **Jacqueline** en **Boudewijn**, ik weet hoe trots jullie op me zijn en ben daar heel dankbaar voor!

Tanya, jij hebt me door de zware periodes gesleept. Je wist als geen ander m'n frustraties en stress te relativeren en daardoor is mijn hele PhD ervaring een stuk gemakkelijker geworden. Je motivatie, stimulatie en constante interesse in m'n werk is goud waard. Ik ben super blij dat ik je heb leren kennen en weet zeker dat samen met jou m'n leven elke dag nog leuker wordt.

About the cover

The cover shows the output of a circle packing algorithm. This algorithm packs circles of random sizes together and then gives the circle the colour of the center pixel of the original image that was used as a template. To me, the circles represent molecules in an artistic way and it simultaneously shows the potential of computers for solving (boring) puzzles in less than a second, something that would take a human a very long time.

Appendix C

Nederlandse samenvatting

Eiwitproductie in cellen wordt beïnvloed door veel regulerende elementen, zowel op DNA- als op RNA-niveau. Een cel gebruikt al deze regulerende elementen om het juiste eiwit, in de juiste hoeveelheid, op het juiste moment te produceren. Deze regulerende elementen kunnen worden gemodificeerd om de cel te dwingen een bepaald eiwit in grote hoeveelheid te produceren. Dit kan een natief of een heteroloog eiwit zijn dat gebruikt kan worden in industrie, voeding of medicijnen. Dit proefschrift richt zich op translationele regulerende elementen ofwel de elementen die bepalen hoeveel eiwit er wordt geproduceerd vanaf een enkel RNA molecuul. Deze elementen kunnen worden onderverdeeld in de drie delen waaruit een gen bestaat: de 5′UTR sequentie, de eiwit coderende sequentie en de 3′UTR sequentie. Deze sequenties zijn bestudeerd om de regulerende mechanismen beter te begrijpen en daardoor controle over deze sequenties te krijgen zodat ze veranderd kunnen worden in hun meest optimale variant.

Allereerst wordt de relevantie van eiwitproductie, zowel in een industriële als academische omgeving, besproken in **Hoofdstuk 1**. Er wordt een overzicht gegeven van de transcriptionele en translationele kenmerken die bijdragen aan de algehele eiwitproductie. Verder wordt het belangrijkste probleem besproken waarom het optimaliseren van deze kenmerken voor een hoge eiwitproductie zo complex is. De reden voor deze complexiteit is dat veel van deze kenmerken met elkaar samenhangen. Dit betekent dat als de genetische code van een regulerend element wordt gewijzigd om dat element te optimaliseren, de prestatie van andere elementen tegelijkertijd beïnvloed worden. In sommige gevallen kunnen deze onbedoelde bijeffecten zo extreem zijn dat de bedoelde optimalisatie teniet wordt gedaan. Deze complexiteit is de reden waarom in dit proefschrift de focus niet alleen ligt op fundamentele ontdekkingen, maar ook op de ontwikkeling van een generieke praktische benadering om functionele eiwitproductie te optimaliseren. Ten slotte wordt in dit hoofdstuk een overzicht gegeven van verschillende, veelgebruikte eiwitproductiesystemen met hun voor- en nadelen.

Hoofdstuk 2 geeft een overzicht van de vele recente onderzoeken die zijn uitgevoerd om eiwitproductie te optimaliseren. Ontwikkelingen in het genereren en analyseren van grote data sets dragen bij tot een grondiger begrip van veel van de betrokken factoren. Met name bij eukaryoten is een sterk verband gevonden tussen de snelheid van translatie en mRNA-stabiliteit. Een transcript met een hoge translatie snelheid is stabieler omdat, vermoedelijk, de hoge dichtheid van snel bewegende ribosomen het transcript beschermt tegen RNA-afbrekende enzymen. Verder gebruiken studies die zich richten op de initiatie van translatie nieuwe methodes om op een experimentele wijze RNA-secundaire structuren *in vivo* te kunnen bepalen, in tegenstelling tot de momenteel veelgebruikte *in silico* voorspellingen. Deze ontwikkelingen zullen bijdragen aan het begrijpen van de (goede en slechte) impact van secundaire structuren op het translationele proces. Een algemene conclusie is dat de effecten van minder invloedrijke genetische kenmerken moeilijk te onderscheiden zijn, omdat deze effecten vaak worden overschaduwd door andere kenmerken die mogelijk sterker zijn en tegelijkertijd kunnen veranderen. Een goed voorbeeld hiervan is het effect van codongebruik op het translatie-elongatie proces, dat voornamelijk wordt overschaduwd door veranderingen in translatie-initiatie veroorzaakt door veranderingen in secundaire structuren. Machine-learning benaderingen kunnen een oplossing bieden om deze meer genuanceerde, subtiele effecten te onthullen. Het goede nieuws is dat dit kan leiden tot een betere voorspelbaarheid van DNA sequentie kenmerken. Het slechte nieuws is dat deze benaderingen mogelijk niet kunnen leiden tot een groter biologisch begrip vanwege het "black box"-karakter van deze algoritmes.

In **Hoofdstuk 3** werd een genetisch ontwerp bestudeerd dat de effecten van secundaire structuur op de translatie-initiatie kan verminderen, wat een belangrijke beperkende factor is in het algehele translationele proces. Een bicistronisch ontwerp is een natuurlijk voorkomend element waarin de coderende regio's van twee genen elkaar overlappen. Deze overlap veroorzaakt waarschijnlijk een vermindering van secundaire mRNA-structuren rondom de Shine-Dalgarnosequentie van het tweede gen (wat codeert voor het eiwit dat geproduceerd moet worden). De verklaring van dit fenomeen is dat het ribosoom dat het eerste gen vertaalt ontvouwing-activiteit vertoont en daardoor eventuele structuren ontvouwd rondom de Shine-Dalgarnosequentie van het tweede gen. Dit bevorderd de translatie-initiatie frequentie van het tweede gen. Het bicistronisch ontwerp maakt het mogelijk om de relatief subtiele bijdrage van codongebruik te bestuderen (Hoofdstuk 4), omdat effecten van secundaire structuren waarbij de Shine-Dalgarnosequentie betrokken is vaak de effecten van codongebruik op translatie-elongatie overschaduwen (zoals besproken in Hoofdstuk 1 en 2). Om de potentie van dit genetisch ontwerp te testen, werden 11 codon sequenties tot expressie gebracht, die gegenereerd waren met verschillende codon-optimalisatie-algoritmes, met en zonder een bicistronisch ontwerp. We hebben waargenomen dat een bicistronisch ontwerp het expressie niveau drastisch kan verbeteren en dat het de relatieve prestaties van verschillende codon-optimalisatie-algoritmes verandert. We toonden verder aan dat het bicistronische ontwerp het potentieel heeft om de expressie van constructen die werden beperkt door een sterke secundaire structuur volledig te redden. We concludeerden dat de integratie van een bicistronisch ontwerp zeer waardevol is in algemene expressie vectoren omdat ze de eiwitproductieniveaus consequent verbeteren. Ten slotte, bij het vergelijken van verschillende codon-optimalisatie-algoritmen of het analyseren van codongebruik, zal het gebruik van een bicistronisch ontwerp het effect van codongebruik op translatie-elongatie versterken. Dit geeft een meer eerlijke vergelijking omdat de effecten minder afhankelijk zijn van secondaire structuren met de 5'UTR-sequentie.

In **Hoofdstuk 4** is het effect van codongebruik op de totale eiwitproductie onderzocht. Zoals besproken in Hoofdstuk 3, is een bicistronisch ontwerp onmisbaar in codon studies als de effecten van translatie-elongatie bestudeerd worden. We hebben daarom een bicistronisch ontwerp opgenomen in al onze gegenereerde codonsequenties. Om de effecten van codons op translatie te observeren, hebben we gekozen voor synonieme codon-randomisatie van de volledige coderende regio van het gen in tegenstelling tot een kleiner specifiek gebied, zoals vaak werd gedaan in eerdere onderzoeken. We hebben een enorm aantal verschillende codonsequenties (350.000) gegenereerd die allemaal coderen voor een rood-fluorescerend eiwit (mRFP) maar met verschillende expressie niveaus. Een subset van 1459 codon sequenties werd geselecteerd die het volledige expressiebereik bestrijkt. Goed presterende sequenties bleken de veelgebruikte (commerciële) en recent voorgestelde (academische) *in silico* codon-optimalisatie algoritmes te overtreffen. Dit laat zien dat de randomisatie benadering potentie heeft als alternatief voor een codon-optimalisatie-algoritme, zeker als dit gekoppeld kan worden aan een automatische selectie strategie (zie Hoofdstuk 6). Verder zijn twee verschillende machine learning-algoritmen getraind op onze dataset en voor ons best presterende algoritme hebben we een Pearson-correlatie van 0,803 verkregen. Daarna hebben we gelimiteerde informatie gebruikt in het machine learning-algoritme zodat we konden achterhalen in welk gedeelte van de codon sequentie de meeste voorspellende informatie zit. Dit deden we door bijvoorbeeld alleen informatie over de eerste 20 codons mee te geven aan het algoritme en vervolgens te kijken naar de voorspelbaarheid van het algoritme. In wezen maakt deze benadering het mogelijk om de beperkende factor in het translationele proces te detecteren. Enigszins tot onze verbazing zagen we dat ondanks het gebruik van een bicistronisch ontwerp de meeste expressieverschillen nog steeds konden worden verklaard door het codongebruik van de eerste 9 aminozuren en door de secundaire structuurformaties tussen de eerste 9 codons en de 5′UTR. Codongebruik gedurende de rest van de coderende regio had wel invloed op de translatie-efficiëntie, maar in mindere mate. We kunnen nu concluderen, dankzij onze volledige randomisatie-aanpak, dat secundaire structuren rond de 5′UTR en eerste 9 codons, en niet het algehele codongebruik, de primaire determinant is van translatie-efficiëntie en dus van eiwitproductie.

In **Hoofdstuk 5** hebben we een nieuw type van translationele koppeling ontdekt. We hebben waargenomen dat in een operon-ontwerp de translatie-efficiëntie van een tweede gen de translatie van een gen daarvoor kan beïnvloeden. Deze koppeling kan relevante implicaties hebben bij het aanbrengen van wijzigingen (invoegingen, deleties, herschikkingen) binnen operons. Dat betekent ook dat fenotypische effecten niet exclusief kunnen worden toegeschreven aan een gen-knock-out binnen een operon, omdat dat gen zelf ook de expressie van omliggende genen kan beïnvloeden. Verder ontdekten we het substantiële effect dat een transcriptionele terminator kan hebben op het translationele proces (tot wel een 50-voudige toename). De 3′UTR werd in meer detail onderzocht door een gerandomiseerde sequentie van 30 nucleotiden in te voegen tussen het stopcodon en de terminator. Deze regio bleek het potentieel te hebben om de algehele eiwitproductie te beïnvloeden. Interessant is dat, in tegenstelling tot 5′UTR-sequenties, deze sequentie onafhankelijk van de genomische context werkte. Een goed presterende 3′UTR-sequentie leidt tot hoge expressie, ongeacht de eiwitcoderende sequentie waar het achter wordt geplaatst. De intrigerende mogelijkheid bestaat dat dit een generiek controlesysteem is (althans bij bacteriën), hoewel dit nog moet worden aangetoond. Wat we in dit stadium wel weten, is dat deze regio nuttig is als betrouwbare afstelling van genexpressie in *E. coli*.

Tot slot, bediscussieer ik al mijn resultaten in **Hoofdstuk 6**. Verder worden er initiële resultaten gepresenteerd die mogelijk kunnen leiden tot een selectie systeem voor hoge eiwitproductie in bacteriën. Het niveau van eiwitexpressie werd gekoppeld aan de productie van een gen dat de bacterie nodig heeft om antibiotica te weerstaan. Hierdoor kan de bacterie alleen overleven als hij voldoende eiwit produceert. Deze methode kan gebruikt worden om uit een groot aantal DNA sequenties de meest optimale voor eiwitproductie te selecteren. Ten slotte presenteer ik een stroomdiagram voor stappen die nu genomen kunnen worden om eiwitproductie te verbeteren.

Appendix D

List of publications

Nieuwkoop, T., Claassens, N. J., & van der Oost, J. (2019). Improved protein production and codon optimization analyses in Escherichia coli by bicistronic design. Microbial biotechnology, 12(1), 173-179. ISSN: 1751-7915. DOI: 10.1111/1751-7915.13332

Nieuwkoop, T., Finger-Bou, M., van der Oost, J., & Claassens, N. J. (2020). The ongoing quest to crack the genetic code for protein production. Molecular cell, 80.2, 193-209. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2020.09.014

Creutzburg, S. C. A.*, **Nieuwkoop, T.***, Zegers, T. & van der Oost, J. (2021) Translational feed-forward and feed-back control. *Manuscript in preparation*

Nieuwkoop, T.*, Terlouw, B. R.*, van der Oost, J. & Claassens, N. J. (2021). Finding determinants of protein translation efficiency via codon randomization and machine learning. *Manuscript in preparation*

Tytgat, H. L. P.*, Swarts, D. C.*, Verver, G. M., Elzinga, J., **Nieuwkoop, T.**, Huuskonen, L., Reunanen, J., Matamoros-Sanchez, A. Z., Forgoine, R. E., Martín-Santamaría, S., van der Oost, J. & de Vos, W. M. (2021). The Akkermansia muciniphila Amuc_1100 Protein Is A Gut-Stable Functional Dimer Associated with Type IV Pili Production and Signaling to the TLR2 Receptor. *Manuscript in preparation*

*equal contribution

Patent applications

T. Nieuwkoop, S. C. A. Creutzburg and J. van der Oost. 3′UTR (Filed in October, 2020)

T. Nieuwkoop, N. J. Claassens and J. van der Oost. Codon random genes (filed in April, 2021)

Appendix E

Overview of completed training activities

Discipline specific activities

- Microbiology Centennial Symposium, Laboratory of Microbiology and Wageningen University & Research, Wageningen (NL) (2017) **
- BaSyC WP5 meeting, BaSyC, Wageningen (NL) (2018, 2020)*
- BaSyC "kick-start" training program, BaSyC, Delft; Amsterdam; Groningen; Wageningen; Nijmegen (NL) (2018)*
- BaSyC International Symposium, BaSyC, Delft (NL) (2018)**
- GASB II Symposium, German Association for Synthetic Biology, Berlin (DE), (2018)**
- BaSyC spring meeting, BaSyC, Delft (NL), (2019, 2020)
- Sound of Biotech Symposium, Nederlandse Biotechnologie Vereniging, Ede (NL), (2019)**
- Protein Synthesis and Translational Control Symposium, EMBO, Heidelberg (DE), (2019)**
- BaSyC International Symposium, BaSyC, Texel (NL), (2020)

*oral presentation, **poster presentation

General courses

- PhD Workshoop Carousel, WGS, Wageningen (NL), (2018)
- Scientific Writing, WGS, Wageningen (NL), (2018)
- Applied Statistics, VLAG, Wageningen (NL), (2018)
- Career orientation, WGS, Wageningen/Online (NL), (2020)
- Biobusiness summerschool, Hyphen Projects, Online (NL), (2020)

Other activities

- Preparation of research proposal (2017)
- iGem supervision (2018)
- PhD trip MIB-SSB, Boston; New York (US) (2019)
- Bacterial Genetics group meetings
- PhD meetings Laboratory of Microbiology

Appendix F

About the author

Thijs Nieuwkoop was born on the 9th of April 1992 in Eindhoven, The Netherlands. After graduating from the Pax Christi College in 2011, he started his BSc in Biotechnology at Wageningen University. His bachelor's thesis was performed at the department of Virology under the supervision of Prof. Dr. Monique van Oers. It encompassed research on prolonging Sf9 cell viability by over-expressing eIF4E for increased protein production in the baculovirus expression system.

In 2014 he continued with a Masters in Biotechnology at Wageningen University with a specialization in Medical Biotechnology. His master's thesis was performed at the department of Microbiology under the supervision of Prof. Dr. John van der Oost. Here he started his research on the codon bias phenomenon and discovered new translational coupling mechanisms in operons. Additionally, the remarkable effect of a strong transcriptional terminator on gene expression was outlined. In 2016 he did at internship at Synthon Biopharmaceuticals in Nijmegen, studying the effect of novel codon optimization methods on the production of monoclonal antibodies in human cells and CHO cells.

In 2017 he started a PhD at the Laboratory of Microbiology at Wageningen University under the supervision of Prof. Dr. John van der Oost and Dr. Nico Claassens. Here he continued his work on codon bias and additional genetic features that contribute to protein production. He was also part of the BaSyC consortium whose aim is to develop a synthetic cell.

The research described in this thesis was financially supported by
the Innovation Programme Microbiology (Wageningen University), and
The Dutch Research Council (NWO)
Gravitation Project BaSyC (024.003.019) to Prof. Dr John van der Oost.

Financial support from Wageningen University
for printing this thesis is gratefully acknowledged.

Printed by proefschriftmaken.nl | | https://www.proefschriftmaken.nl