NTNU – Trondheim
Norwegian University of
Science and Technology

Recombinant Gene Expression in *Escherichia coli*

An Experiment and Literature-Based Study of
the Roles of the 5'-Ends of Target Genes

Jon Andreas Lorentzen

Biotechnology (5 year)
Submission date: October 2012
Supervisor: Svein Valla, IBT
Co-supervisor: Simone Balzer, IBT

Norwegian University of Science and Technology
Department of Biotechnology

Preface

This thesis is submitted as part of the requirements for receiving the degree of Master of Science (M.Sc) in Biotechnology at the Norwegian University of Science and Technology (NTNU), Department of Biotechnology. The work presented has been carried out under the supervision of Professor Svein Valla and Phd candidate Simone Balzer.

This master thesis differs from most others since it consists of two separate parts. The topics of the two parts overlap to some degree, but are otherwise independent of each other. The reason for this solution is because of an unforeseen interruption of the laboratory work due to injuries and health issues.

The first part is the one most resembling a standard master thesis, and is based on the laboratory work performed. It should be noted that much of the results presented were actually thought of as initial research leading to further experiments. Part two was added to compensate for the somewhat shortened laboratory work and is a review article covering recent research, and is a result of an extensive literature search.

Trondheim, 8th October 2012

Jon Andreas Lorentzen

Acknowledgements

My time as a master student has surely been a bumpy one. At some times it felt like fate did not want me to end the good life as a student. I started off by breaking my elbow, but returned fairly quickly to the lab after an operation. Then I ruptured my Achilles tendon helping me rebuild my elbow strength by walking around on crutches for two months. Not too long after I had returned to the lab I became sick and was rendered more or less useless for three months. It is now known that I have a genetic mutations in the gene associated with Marfan syndrome, which I find somewhat ironic after studying genes for five years. However, I have never lost my good spirit much due to the good people around me.

I have to thank my supervisors and the department of Biotechnology, who have met me with nothing but understanding and helped with finding a solution allowing me to finish my master thesis. Special thanks go to my two supervises: Simone Balzer, helping me on a day to day basis and for widely improving my competence in the lab, and Svein Valla for sharing from his vast knowledge and experience. I would like to thank my family for support and my mother especially for helping me in the search for grammatical errors. Finally, I will give one million thanks to my girlfriend Anne, for taking care of me at my less functional times and for lightening my days when I have barricaded myself in my room working on this thesis.

Part One: The Experimental Study

5’-UTR-Optimization for Maximal Enhancement of Transcription and Translation in *Escherichia coli*

Abstract: The 5’-untranslated region (5’-UTR), and the DNA region corresponding to it, have been shown to have a significant influence on the expression of genes both at transcriptional and translational level. Since transcription and translation are two independent mechanisms a 5’-UTR sequence will probably not be optimal for both. The suggested solution presented in this study is to design a long 5’-UTR composed of one transcription stimulating and one translation stimulating region. The stimulating regions consist of 5’-UTR variants independently identified as transcription or translation stimulating by screening for the desired trait, as well as translation stimulating 5’-UTR variants designed using a bioinformatics tool.

The results indicated that 5’-UTR fusions tend to introduce limiting factors yielding a reduced gene expression. However, some 5’-UTR fusions successfully resulted in high gene expression and one variant surpassed both of its components showing a possible additive effect of stimulating both transcription and translation in the form of 5’-UTR fusions. This indicates that testing a relatively small number of different sequences gives a good chance of success. The method also proved viable to increase the expression of low expressive 5’-UTR variants while maintaining low uninduced expression. In addition 5’-UTR fusions containing *in silico* designed translation stimulating regions have the potential of reaching expression levels on par with the levels reached by fusions containing 5’-UTR variants identified through screening.

Sammendrag: Den 5'-ikke-kodende regionen (5'-UTR), og DNA regionen som tilsvarer denne, har vist seg å ha en betydelig innflytelse på uttrykk av gener både på transkripsjons og translasjons nivå. Siden transkripsjon og translasjon er to uavhengige mekanismer vil en 5'-UTR sekvens trolig ikke være optimal for begge. Løsningen som legges fram i denne oppgaven er å designe en lang 5'-UTR som består av en transkripsjons-stimulerende og en translasjons-stimulerende region. De stimulerende regionene består av 5'-UTR varianter som tidligere har blitt identifisert som enten transkripsjons- eller translasjons-stimulerende ved seleksjon for den ønskede egenskapen, i tillegg til translasjons-stimulerende 5'-UTR varianter designet ved bruk av et bioinformatisk verktøy.

Resultatene indikerer at 5'-UTR fusjonene har en tendens til å innføre begrensende faktorer som fører til redusert genuttrykk. Noen 5'-UTR fusjoner resulterte imidlertid i høyt genuttrykk og en variant førte til høyere uttrykk enn hva bestanddelene gjorde alene, som tyder på en mulig additiv effekt ved å stimulere både transkripsjon og translasjon i form av en 5'-UTR fusjon. Dette indikerer at å teste et relativt lite utvalg av 5'UTR-sekvenser gir en god sjanse for suksess. Metoden viste også potensial til å øke ekspresjonen til lav-uttrykks varianter og samtidig beholde det lave uinduserte genuttrykket. I tillegg viste det seg at 5'-UTR fusjoner med *in silico* designede translasjons-stimulerende regioner kan oppnå ekspresjonsnivå på linje med nivåene oppnådd av 5'-UTR fusjoner bestående av regioner identifisert ved seleksjon.

Table of content:

Abbreviations

BSA - Bovine serum albumin

bla – Gene encoding β-lactamase

bp – base pair(s)

celB- Gene encoding phophoglucomutase

DNA - Deoxyribonucleic acid

DTT – Dithiothreitol

GTP – Guanine triphosphate

mRNA – Messenger ribonucleic acid

nm – nanometer

ng – nanogram

OD - Optical density

ON - Over night

Ori - Origin of replication

RBS – Ribosome binding site

RNA - Ribonucleic acid

RNAP – RNA polymerase

RNAse- Ribonuclease

rpm - Revolutions per minute

rRNA - Ribosomal ribonucleic acid

RT - PCR Real-time PCR

SD - Shine-Dalgarno

TOL plasmid - Toluene-degradative plasmid

TSS – Transcription start site

tRNA - Transfer-ribonucleic acid

UTR - Untranslated region

1 Introduction

1.1 Recombinant gene expression in *E. coli*

Gene expression in an organism is extensively regulated to best ensure its viability. In order to achieve effective production of a gene of choice in a host organism, it is necessary to manipulate the regulatory mechanisms of gene expression. This section will cover some of the fundamental mechanisms in regulation of gene expression on the pathway from DNA to functional protein and commonly used strategies for achieving effective recombinant production.

1.1.1 *E. coli* as a production host

E. coli was the first host organism utilized for the recombinant production of therapeutic proteins, and it has maintained its prominent position as a production host being responsible for nearly 30% of the recombinant therapeutic proteins produced today (Huang et al. 2012). The advantages with *E. coli* are favourable industrial qualities, like low-cost medium requirements and rapid growth rate, as well as extensive knowledge of the organism and vast experience on it as a production host. Drawbacks are its limited ability to perform certain posttranslational modifications, protein maturation and disulfide formations, which is part of the reason several other host options have arisen, including yeasts and mammalian cells (Huang et al. 2012).

1.1.2 Regulation at the transcriptional level

Transcription is the step where RNA is synthesized on the basis of a DNA template, and it consists of three phases; initiation, elongation and termination. The central DNA element responsible for the initiation of transcription is the promoter element located upstream of the gene. The promoter element consists of four regions: the two hexamers named the -10 and -35 region after their location in respect to the transcriptional start site (TSS), the spacer sequence between them, and the A/T rich UP element (fig. 1-1)(Saecker et al. 2011).

Figure 1-1: A σ^{70} Promoter element, showing the consensus sequence of the -10 and -35 region recognized by the σ factor.

The RNA synthesis is carried out by the multi subunit RNA polymerase (RNAP) enzyme. The RNAP consists of five subunits with the stoichiometry $\alpha_2\beta\beta'\omega\sigma$ where the $\alpha_2\beta\beta'\omega$ part form the core enzyme region which is able to synthesize RNA from a DNA template, and the *σ*-factor which is responsible for recruiting the RNAP to the promoter element. The -10 and -35 regions are the most essential for binding of the σ-factor and the closer they are to the consensus sequence (fig. 1-1) the more effective is the binding of the RNAP to the promoter. However, a promoter which is at perfect consensus for all the regions would bind RNAP too tightly (Browning et al. 2004). Several different σ-factors are produced in *E. coli* recognizing different promoter elements. The most common is σ^{70}, but other σ-factors come to play during certain stress situations and competes with σ^{70} for binding to the RNAP complex (Saecker et al. 2011). During transcription initiation RNAP must first interact with the promoter sequence in such a way that the double stranded DNA is unwound around the TSS, forming what is known as the "open complex". Then the formation of the first phospho-diester bonds in the nascent RNA sequence is established, leading RNAP into the elongation complex synthesizing the rest of the RNA strand (Browning et al. 2004). It has been shown that the RNAP might begin to synthesize RNA several times before it escapes the promoter and enters the elongation phase. This phenomenon is termed abortive initiation and leads to the synthesis of short aborted transcripts (Goldman et al. 2009). After promoter escape the RNAP enters the elongation complex, which can exist in several conformational states regulating the efficiency of elongation as a response to different interactions (Erie 2002). Transcription termination is executed through two mechanisms. Intrinsic termination takes place at certain template palindromic sequences, forming stable hairpin structures in the newly synthesized RNA, followed by a T-stretch. The combination of the hairpin followed by unpaired U residues destabilizes the RNAP, leading to transcript termination (Nudler et al. 2002). The other mechanism is factor dependent termination, which depends on factors like the Rho protein. The Rho protein complex binds accessible RNA, pulling the transcript away from RNAP (Nudler et al. 2002).

Most regulatory mechanisms target the transcription initiation step, due to the disadvantage of wasting unnecessary energy by aborting a process mid-way (Browning et al. 2004). The supply of RNAP is limited in a cell. Hence, the promoter sequence, which determines the efficiency of RNAP binding, is a crucial factor in determining the transcript amount produced. Trans-acting factors may also contribute both by activating and repressing transcription (Browning et al. 2004). In addition to the importance of the promoter sequence it has been shown that the DNA sequence downstream of the promoter, corresponding to the 5'-untranslated region (5'-UTR), can have a significant influence on gene expression. This effect is hypothesized to be linked with promoter escape and/or the rate of formation of the open complex (Berg et al. 2009).

1.1.3 Properties of an expression vector

The transcript amount produced from a gene located in a vector plasmid is highly affected by the plasmid copy number and the promoter controlling the gene. The plasmid copy number is determined by the origin of replication (*ori*), and a high copy number is associated with high production and plasmid stability. However, a large number of plasmids might pose a significantly metabolic burden on the cell. A selection marker (usually antibiotic resistance) to select for cells harbouring the plasmid is also essential (Jana et al. 2005). In addition to choosing a strong promoter it is advantageous to choose a tightly regulated promoter, which will lower the burden on the cells when in the growth phase. Simple and cost-effective induction is also important, especially at industrial scale (Jana et al. 2005).

1.1.4 mRNA decay

In addition to the transcription efficiency the transcript amount in the cell is regulated by the speed at which the mRNA is degraded (mRNA decay rate). The RNA degradation is mediated by RNases which can operate alone or as part of degradation complexes (Burger et al. 2011). The most prominent RNase in *E. coli* is RNase E, catalyzing the rate limiting step of cleaving single stranded A/U rich regions leaving fragments vulnerable for further degradation. In addition to RNases cutting RNA internally (endonucleases) like RNase E there is RNases cleaving RNA from the terminal ends (exonucleases) either in the 5'→3' direction or the 3'→5' direction (Arraiano et al. 2010). Several factors will influence the mRNA decay rate. Secondary structures can make the mRNA less accessible for degradation. Sequences

containing poly-A stretches are more susceptible for degradation since poly-A stretches are the preferred binding site for many RNases. The translation rate will affect the mRNA decay rate since ribosomes covering it will protect it from degradation. In addition trans-acting factors may make sequence stretches preferred by RNases more or less accessible (Arraiano et al. 2010).

1.1.5 Regulation at the translational level

Translation is the synthesis of polypeptides by ribosomes using an mRNA sequence as template. As transcription, it consists of the three steps; initiation, elongation and termination. The rate limiting step of translation is the initiation, which begins with the small ribosomal subunit (30S) interacting with three initiation factors (IF 1,2 and 3) favouring the binding of mRNA and the initiator tRNA (fMet-tRNA^{fMet}) (Simonetti et al. 2009). 30S binds the mRNA at the Shine Dalgarno (SD) sequence with consensus GGAGG (in *E. coli)* located in the 5'-UTR by base pairing with the 16S rRNA 3'-end (anti SD), and the initiator tRNA binds the start codon. Then the large ribosomal subunit (50S) is recruited followed by the release of the initiation factors and entrance to the elongation phase (Simonetti et al. 2009).

The mRNA nucleotides are interpreted in series of three, named codons. Each codon translates to an amino acid (except for the stop codons). Since there are more possible codon combinations than amino acids, different codons may encode the same amino acid (Taylor et al. 1989). tRNAs bearing amino acids are brought to the ribosome as a complex with the elongation factor Tu and GTP. The amino acid is added to the peptide chain using the energy provided by dephosphorylating the GTP (Kaczanowska et al. 2007). Translation termination takes place when the ribosome arrives at one of the three possible stop codons in the mRNA sequence, which is recognized by a release factor triggering the release of the peptide chain (Kaczanowska et al. 2007).

Much of the regulation of the translation efficiency affects the translation initiation. To what extent the SD sequence matches the consensus and also the distance of the SD to the start codon will affect the translation rate (Chen et al. 1994; Ma et al. 2002). The remaining 5'-UTR sequence in addition to the SD sequence has also been found to have a significant influence on the translation efficiency (Berg et al. 2009; Berg et al. 2012). The most effective and abundant start codon is AUG (Laursen et al. 2005), but alternative start codons can be found in the *E. coli* genome (Blattner et al. 1997). In addition, factors influencing the

availability of the SD sequence and/or the start codon will also regulate translation initiation. This can be trans-acting elements like sRNA (small RNA) (Storz et al. 2011), mRNA binding proteins (Babitzke et al. 2009), and RNA structures changing conformation due to interacting with small molecules (riboswitches) (Breaker 2011) or through temperature change (Kortmann et al. 2012). Moreover Secondary structures are also known to interfere with the translation initiation (Kudla et al. 2009).

For optimal translation one should have a strong SD sequence at an optimal distance to the start codon (Chen et al. 1994; Vimberg et al. 2007). Another important factor is that many heterologous proteins will have strong secondary structures at the 5'-UTR and the initial coding sequence of the mRNA compared to endogenous genes, reducing the translation initiation rate (Tuller et al. 2010b). Hence reducing the 5'-end stability have often proven to increase recombinant expression (Cèbe et al. 2006; Kudla et al. 2009; Jung et al. 2010).

A common tactic of achieving high expression levels of heterologous genes is codon optimization. Several codons codes for the same amino acid, but the abundance of each codon may differ among organisms. Thus, a gene harbouring many rare codons can hamper the translation (Huang et al. 2012). A solution is to synthetically make a new gene sequence with better suited codons, designed by a bioinformatics tool made for that purpose (Villalobos et al. 2006). There are several methods in which the optimal codon composition can be selected. The simplest approach is to use the most common codons in the host organism by optimizing the Codon Adaptation Index (CAI) score (Sharp et al. 1987). Alternatively the tRNA Adaptation Index (tAI) can be used, which is a measure for the tRNA usage based on that the tRNA gene copy number is correlated with the tRNA abundance and the codon preference of the organism (dos Reis et al. 2003). Traditional codon optimization has been challenged in recent years by the research of Welch et al. who indicated that picking the codons recognized by the tRNAs which are most efficiently charged during amino acid starvation is the most effective strategy (Welch et al. 2009). In addition, Tuller et al. states that ineffective codons within the 30-50 first codons might by advantageous due to reduced ribosomal jamming (Tuller et al. 2010a).

1.1.6 Protein folding and degeneration

After a polypeptide is synthesized through translation it needs to fold into the correct tertiary structure to become a functional protein. The conditions in a cell will often disfavour

spontaneous correct folding, for instance due to high protein concentration which may lead to protein aggregation (Dougan et al. 2002). As a consequence the cell harbour proteins called chaperones, which help to create the proper environment for correct folding. Certain chaperones may come to play during acute stress situations like heat shock (Walter et al. 2002). Protein degradation by proteases will also affect the gene expression (Dougan et al. 2002), and the extent of protein degradation may be regulated as a response to environmental factors, like starvation (Schweder et al. 1996).

1.2 Constructs and sequence elements used in this study

The expression vectors utilized in this study are derived from RK2 based, tightly controlled, broad host range plasmids containing the XylS/*Pm* promoter system (Blatny et al. 1997). These mini RK2 plasmids contain the origin of replication, *oriV*, and the *trfA* gene coding for the replication initiation protein (Perri et al. 1991). The XylS/*Pm* promoter system originates from the TOL plasmid of *Pseudomonas putida*. In this system the XylS protein is constitutively produced from its promoter, but only acts as a transcription activator of the *Pm* promoter when bound by benzoate derivatives, such as m-toluic acid (the inducer used in this study) (fig. 1-2) (Ramos et al. 1997). The XylS/*Pm* system has been shown to produce recombinant proteins at industrial levels in *E. coli (Sletta et al. 2004)*. The plasmids used in this study harbour the *kan* gene, encoding kanamycin resistance, which was used as a selective marker.

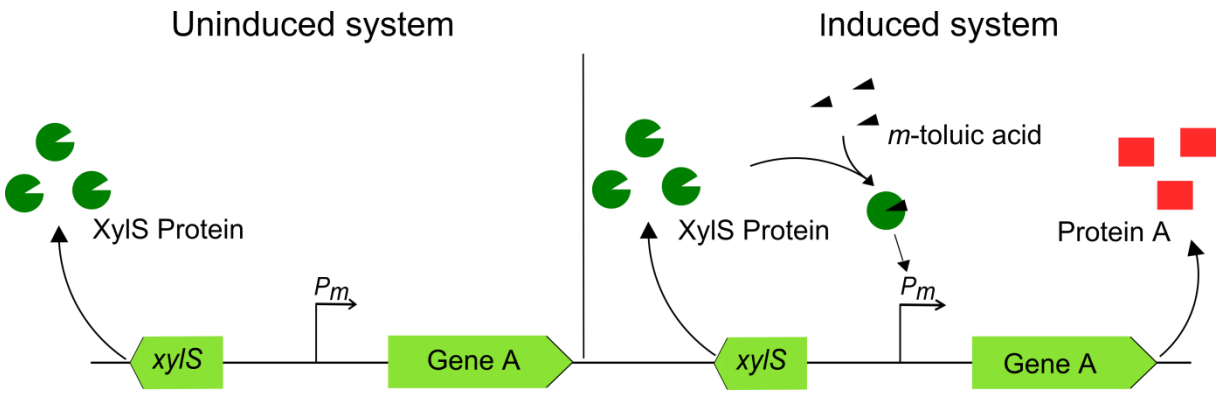

Figure 1-2: the XylS/*Pm* promoter system and how the system is induced by benzoate derivatives (in this case m-toluic acid). XylS proteins are constitutively synthesized, but will only function as an activator of the *Pm* promoter when bound to a benzoate derivate.

The use of the penicillin resistance gene *bla* has previously been used as a reporter gene for the XylS/*Pm* system (Berg et al. 2009). This allows screening for high productive variants by plaiting cells on agar with an increasing ampicillin (a type of penicillin) concentration, since the resistance level correlates with the expression of *bla*. In addition the expression can be further explored by monitoring the rate of penicillin degradation (Winther-Larsen et al. 2000; Berg et al. 2009)(also see section 2.2.3).

The pIB11 plasmid (fig. 2-1) constructed by Berg et al. is a plasmid with the properties described above used to research the effect of the 5'-UTR on gene expression (Berg et al. 2009; Lale et al. 2011; Berg et al. 2012), and it is also utilized in this study.

Another RK2 based plasmid utilizing the XylS/*Pm* system used in this study is the pAO-double construct (fig. 2-2) designed for selection of 5'-UTRs enhancing either transcription or translation (Lale, unpublished). In this construct the *Pm* promoter transcribes an mRNA consisting of two genes; *celB* located upstream of *bla* separated by a spacer region (fig. 1-3). The two genes are individually translated, each harbouring its own 5'-UTR. The length of the spacer region between *celB* and the 5'-UTR was ensured to be long enough so that ribosomes translating *celB* would leave the mRNA before reaching the *bla* sequence. Because of this design the 5'-UTR in front of *celB* will not affect the translation of *bla*. Identifying 5'-UTRs stimulating transcription can be done by altering the *celB* 5'-UTR and screen for increased *bla* expression (increased ampicillin tolerance). Any increase in *bla* expression observed will then be a result of an increase in *celB-bla* transcript and not *bla* translation since the *bla* 5'-UTR is unchanged. Identifying translation stimulating 5'-UTRs can be done by altering the *bla* 5'UTR and screen for increased *bla* expression. This will leave the *celB-bla* transcript level unchanged and any increase in *bla* expression will be due to increased translation efficiency.

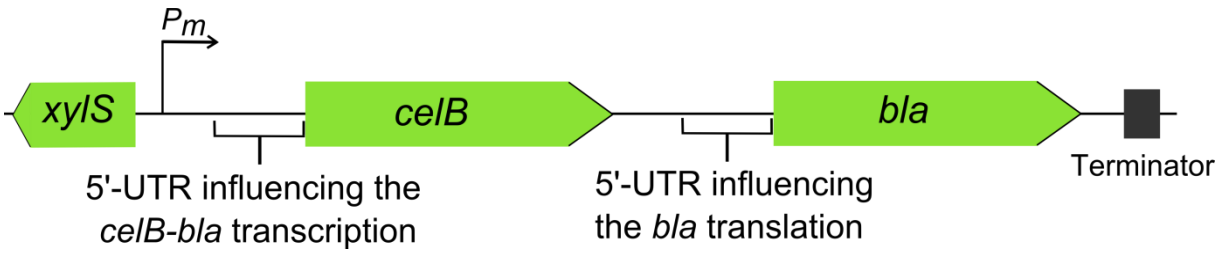

Figure 1-3: Part of the pAO double construct (fig. 2-2) with two individual 5'-UTRs influencing the *bla* expression. The 5'-UTR in front of *celB* will affect the *bla* expression by influencing the *celB-bla* mRNA transcription and the 5'-UTR in front of *bla* will affect the expression by influencing the *bla* translation.

Prior to this study the method described above was utilized to identify transcription and translation stimulating 5'-UTRs from 5'-UTR libraries (Balzer, unpublished). The libraries were made using the same method as in the published 5'-UTR study performed by Berg and colleagues (Berg et al. 2009). Some of the 5'-UTRs identified are used in this study (tab. 1-1).

Table 1-1: The 5'-UTR sequences identified as transcription or translation stimulating through selection screening of a 5'-UTR (Balzer, unpublished) library which are used in this study. The restriction sites (underscored) are PciI at the 5'-end and NdeI at the 3'-end. The SD-sequence is marked in bold, and the mutations in respect to the wt 5'-sequence are marked in red.

	Transcription stimulating 5'-UTRs
wt	AACATGT-ACAATAATAAT**GGAG**TCATGAACATATG
Tr28	AACATGT-A-TATAATAAT**GGAG**AAATGAACATATG
Tr31	AACATGTTACCATGATAAT**GGAG**TCATGAACATATG
Tr36	AACATGT-ACAAGTATAAC**GGAG**TAATGAACATATG
Tr50	AACATGTTACAATAATAAC**GGAG**TCATGTACATATG
	Translation stimulating 5'-UTRs
wt	AACATGTACAATAATAAT**GGAG**TCATGAACATATG
Tn2	AACATGTAGTTTAATA-T**GGAG**TCATTAACATATG
Tn15	AACATGTACAACAATAGT**GGAG**TCATTAACATATG
Tn24	AACATGTACATTAATATA**GGAG**TCATCAACATATG
Tn25	AACATGTACAATGATAAT**GGAG**TCATAAACATATG
Tn47	AACATGTAATAAACTAAA**GGAG**TTATGAACATATG
Tn58	AACATGTACATTAACAAA**GGAG**TCATATACATATG
Tn59	AACATGTACTATTAGATA**GGAG**TCATTAACATATG

Some 5'-UTR variants originating from previously published works have also been further explored in this study. These are summarized in table 1-2 and consist of the LII-11 5'-UTR variant shown to highly increase β-lactamase activity and the LV-1 and LV-2 5'-UTR variants shown to highly increase the *bla* transcription amount (Berg et al. 2009), as well as the 5'-UTR down variants DI-3 and DI-8 leading to very low uninduced expression at the cost of reduced induced expression (Lale et al. 2011).

Table 1-2: The 5'-UTR sequences originating from the previously published work of (Berg et al. 2009), and (Lale et al. 2011). The restriction sites (underscored) are PciI at the 5'-end and NdeI at the 3'-end. The SD sequence is marked in bold, and the mutations in respect to the wt sequence are marked in red.

	5'-UTRs from (Berg et al. 2009)
wt	A<u>ACATGT</u>ACAATAATAAT**GGAG**TCATGAA<u>CATATG</u>
LII-11	A<u>ACATGT</u>TACACAATAAT**GGAG**TAATGAA<u>CATATG</u>
LV-1	A<u>ACATGT</u>ACCATTATAAC**GGAG**TAATGAA<u>CATATG</u>
LV-2	A<u>ACATGT</u>ACCATAATACA**GGAG**TTATGAA<u>CATATG</u>
	5'-UTRs from (Lale et al. 2011)
wt	A<u>ACATGT</u>ACAATAATAAT**GGAG**TCATGAA<u>CATATG</u>
DI-3	A<u>ACATGT</u>GGCATAATAAT**GGAG**TTATGCA<u>CATATG</u>
DI-8	A<u>ACATGT</u>CCCATAATAATGGAGTCATGAA<u>CATATG</u>

1.3 Aims of the study

The overall aim of this study is to identify new 5'-UTR sequences stimulating gene expression at a higher level than those available today. The approach taken is built on the discovery that the 5'-UTR (and its corresponding DNA region) can influence both the transcriptional and translational efficiencies (Berg et al. 2009). Transcription and translation are two separate mechanisms, thus there is a good chance that one 5'-UTR sequence will not

be optimal for both transcription and translation. The suggested solution presented in this study is to design a long 5'-UTR, composed of one transcription stimulating and one translation stimulating region placed at the 5'-end and 3'-end of the 5'-UTR respectively (fig. 1-4). The position of the stimulating regions are based on the assumption that it is most important for the part stimulating transcription to be localized close to the promoter region and transcription initiation while the 3'-end will have the SD sequence in a suitable distance to the start codon and take part in translation initiation.

The stimulating regions will consist of 5'-UTRs independently identified as transcription or translation stimulating by screening for the desired trait. It will be tested whether the long 5'-UTR fusion will lead to an increase in gene expression surpassing the expression achieved by its components alone. In addition the performance of translation stimulating 5'-UTRs made by a bioinformatics tool will be compared to those obtained through selection screening.

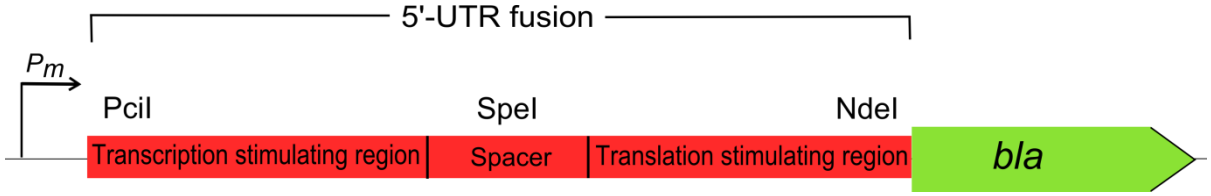

Figure 1-4: An overview of a 5'-UTR fusion consisting of a transcription and a translation stimulating region fused together with a spacer region. The restriction sites located within the 5'-UTR fusion are included in the figure.

5'-UTR sequences with down regulated gene expression and very low uninduced expression levels have been identified (Lale et al. 2011). Some of these will also be combined with a transcription stimulating 5'-UTR to see if the gene expression can increase while maintaining the desired low uninduced expression level.

Some characterization of the short 5'-UTR components used to build up the 5'-UTR fusions will also be included in this study to better understand their transcription and translation stimulating properties.

The 5'-UTR sequence has already shown great potential in increasing recombinant gene expression through both transcription and translation stimulation (Berg et al. 2009; Berg et al. 2012). This study is a step towards finding a 5'-UTR with optimal stimulation of both

transcription and translation, which if successful would represent a new type of approach for the optimization of recombinant gene expression.

1.4 Laboratory experiments and bioinformatics analyses

The laboratory experiments (3.1 and 3.2) performed in this study will fist cover characterization of the transcriptional and translational stimulating traits of some of the previously identified 5'-UTR variants (table 1-1 and 1-2) by utilizing the properties of the pAO double construct (section 1.2, fig. 2-2) as well as protein assay and transcriptional analysis (section 2.2.3). Then the potential of enhancing *bla* expression by combining transcriptional and translational stimulating 5'-UTRs both as separate units in the pAO double construct and as a long 5'-UTR fusion in the pLUTR construct (fig 2-1) will be explored by replica plating (section 2.2.3). The *E. coli* strain utilized throughout this study is DH5α (Tab. 2-1).

The second part (3.3) of this study involves bioinformatics analyses of the 5'-UTR sequences, exploring factors explaining the observed ampicillin tolerance (*bla* expression). The analyses performed cover the translation initiation rate (TIR), the structural energy of the 5'-UTRs, the location of the SD sequence in respect to the secondary structures and the strength of the SD-containing secondary structures.

2 Materials and Methods

2.1 Materials

2.1.1 Recipes

Lysogeny Broth (LB) medium

- 10 g/L Bacto™ Tryptone (Becton, Dickinson& Co.,Sparks, USA)
- 5 g/L Yeast extract (Oxoid, Basingstoke, Hampshire, England)
- 5 g/L NaCl

Components were dissolved in distilled water, then autoclaved at 120 ºC for 20 minutes.

LB Agar (LA)

- LB medium
- 15 g/L DifcoTM Agar Noble (Becton, Dickinson and Company, Sparks, USA)

LB components and agar were dissolved in distilled water, then autoclaved at 120 ºC for 20 minutes. The LA was cooled to~50 ºC, before addition of selective antibiotics and casting of agar plates.

Psi Medium

- 5 g/L Yeast extract (Oxoid, Basingstoke, Hampshire, England)
- 20 g/L Bacto™ Tryptone (Becton, Dickinson& Co.,Sparks, USA)
- 5 g/L $MgSO_4$ (10.24 g/L $MgSO_4$ x $7H_2O$)

The components were dissolved in distilled water and the pH was adjusted to 7.6 using KOH, followed by autoclaving at 120 ºC for 20 minutes.

Transformation Buffer 1 (TFB1)

- 30 mM Potassium acetate (CH_3CO_2K)
- 100 mM RbCl
- 10 mM $CaCl_2$
- 80 mM $MnCl_2$
- 15 % (v/v) glycerol

The components were dissolved in distilled water, and the pH was adjusted to 5.8 using acetic acid followed by sterile filtration (0.22 µm filter). Stored at 4 ºC

Transformation Buffer 2 (TFB2)

- 10 mM MOPS
- 100mM $CaCl_2$
- 10 mM RbCl
- 15% (v/v) glycerol

Components were dissolved in distilled water, and the pH was adjusted to 6.5 using NaOH followed by sterile filtration (0.22 µm filter).

Kanamycin stock solution

- 50 mg/mL Kanamycin sulphate

Kanamycin sulphate was dissolved in distilled water and sterile filtered (0.22 µm filter). Stored at – 20 ºC after aliquotation

Ampicillin stock solution

- 200 mg/mL Ampicillin sodium salt

Ampicillin sodium salt was dissolved in distilled water and sterile filtered (0.22 µm filter) before use the same day.

β-lactamase assay buffer

- 80 mM $K_2H_2PO_4$
- 20 mM NaH_2PO_4

The components were dissolved in distilled water and sterile filtered (0.22 µm filter).

m-toluic acid stock solution

m-toluic acid was dissolved in ethanol making a 1M stock solution.

2.1.2 Plasmids and strains

The bacterial strain and the plasmids used in this study are described in table 2-1.Plasmid maps of the plasmids utilized in this study are shown in figure 2-1 and 2-2.

Table 2-1: Bacterial strains and plasmids used in this project. The selectable marker for all plasmids is Kan^r.

Strain or plasmid	Description	Source or reference
Strain		
Escherichia coli DH5α	General cloning strain. Genotype F^-, *end*A1, *hsd*R17 (r_k^-, m_k^+), *sup*E44, *thi*-1, λ-, *rec*A1, *gyr*A96, *rel*A1, ϕ80d*lacZ*ΔM15.	(Bethesda 1986)
Plasmids		
Single constructs		
PIB11	RK2-based expression vector containing the *xylS/Pm* expression cassette with the penicillin-resistance gene, *bla*, as the reporter gene for *Pm.*	(Berg et al. 2009)
PIB11-LV-1 and LV-2	pIB11 derivate with the LV-1 and LV-2 5'-UTR variant in front of *bla.*	(Berg et al. 2009)
PIB11-DI-3 and DI-8	pIB11 derivate with the DI-3 and DI-8 5'-UTR variant in front of *bla.*	(Lale et al. 2011)
pIB11-Tnx	pIB11 derivates where x indicates the 5'-UTR variant identified by screening for high translation (tn).	(Simone Balzer, unpublished)
pIB11-Trx	pIB11 derivates where x indicates the 5'-UTR variant identified by screening for high transcription (tr).	(Simone Balzer, unpublished)
pLUTR	pIB11 derivate with a longer 5'-UTR in front of *bla* composed of 2 wt 5'-UTRs fused together with a spacer sequence in between (fig 2-1).	This study
pLUTR-x y	pLUTR derivates. The x and y indicate the 5'-UTR sequences the 5'-UTR fusion is composed of, x being the upstream part and y the downstream part of the 5'-UTR fusion.	This study

Double constructs

pAO-Tr	Expression vector containing the *xylS/Pm* expression cassette. The construct contains *Pm* – *celB* – spacer – *bla* with restriction sites making it possible to switch out the 5'-UTR region in front of *celB (fig* 2-2).	(Rahmi Lale, unpublished)
pAO-Tn	Expression vector containing the *xylS/PM* expression cassette. The construct contains *PM* – *celB* – spacer – *bla* with restriction sites making it possible to switch out the 5'-UTR region in front of *bla (fig* 2-2).	(Rahmi Lale, unpublished)
pAO-TrTn	A combination of pAO-Tr and pAO-Tn made when switching out the *bla* fragment of pAO-Tr with the one in pAO-Tn (by digesting with EcoRI and BamHI) fig 2-2.	This study
pAO-Trx	pAO-Tr derivates where x indicates the 5'-UTR variants identified when screening for high transcription (Tr).	(Simone Balzer, unpublished)
pAO-Tnx	pAO-Tn derivate where x indicates the 5'-UTR variant identified when screening for high translation (Tn).	(Simone Balzer, unpublished)
pAO-Trx Tny	pAO-TrTn derivate where x and y indicate the mutated 5'-UTR sequence in front of *celB* and *bla* , respectively.	This study

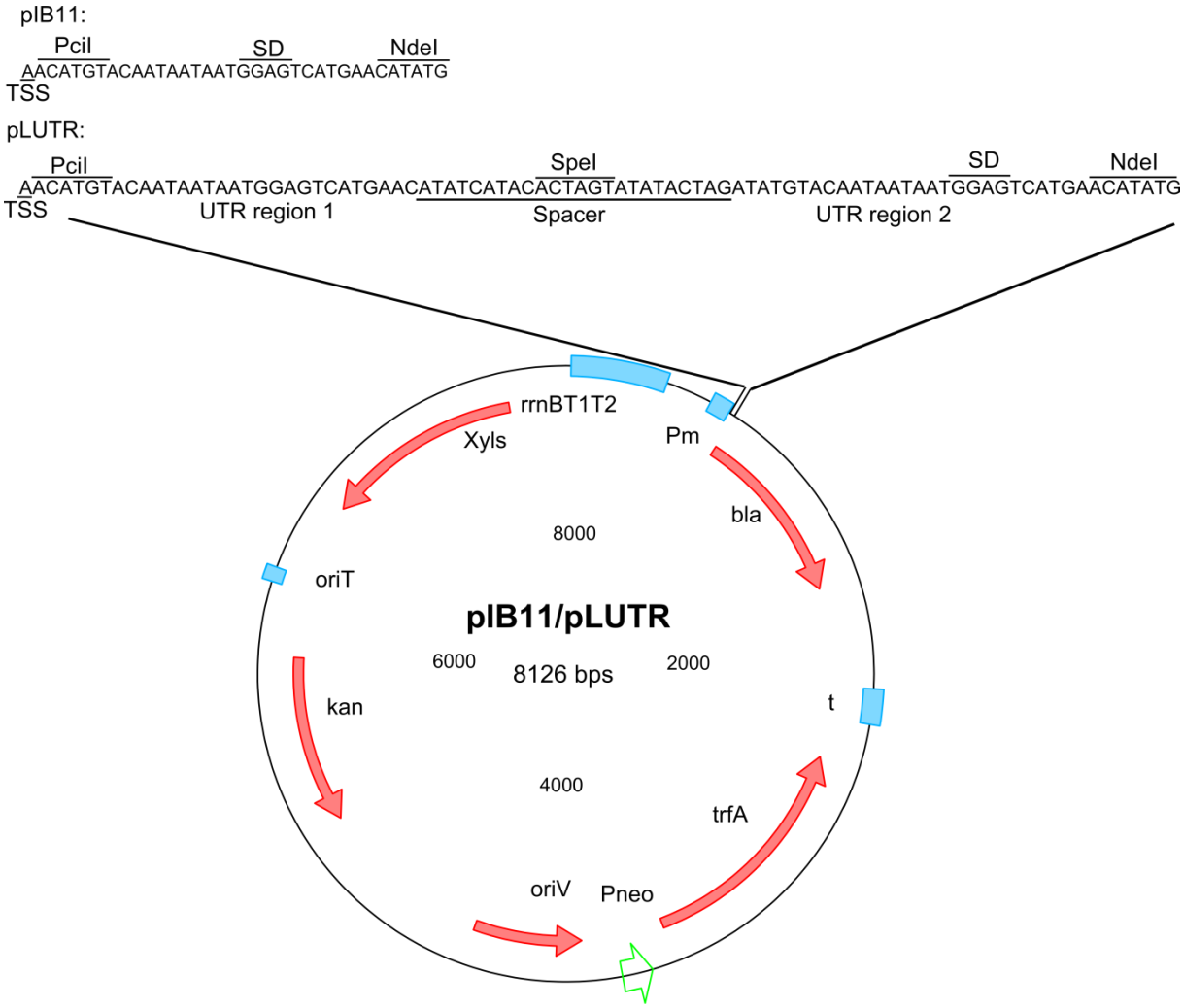

Figure 2-1: Plasmid map of plB11/pLUTR containing the *xylS/Pm* expression cassette and *bla* as reporter gene. Kanamycin is used as a selective antibiotic for the plasmid during culturing. The 5'-UTR sequence in front of *bla* is expanded with annotated restriction sites, Shine-Dalgarno (SD), transcription start site (TSS), *bla* start codon, as well as the spacer and the two 5'-UTRs of the 5'-UTR fusion.

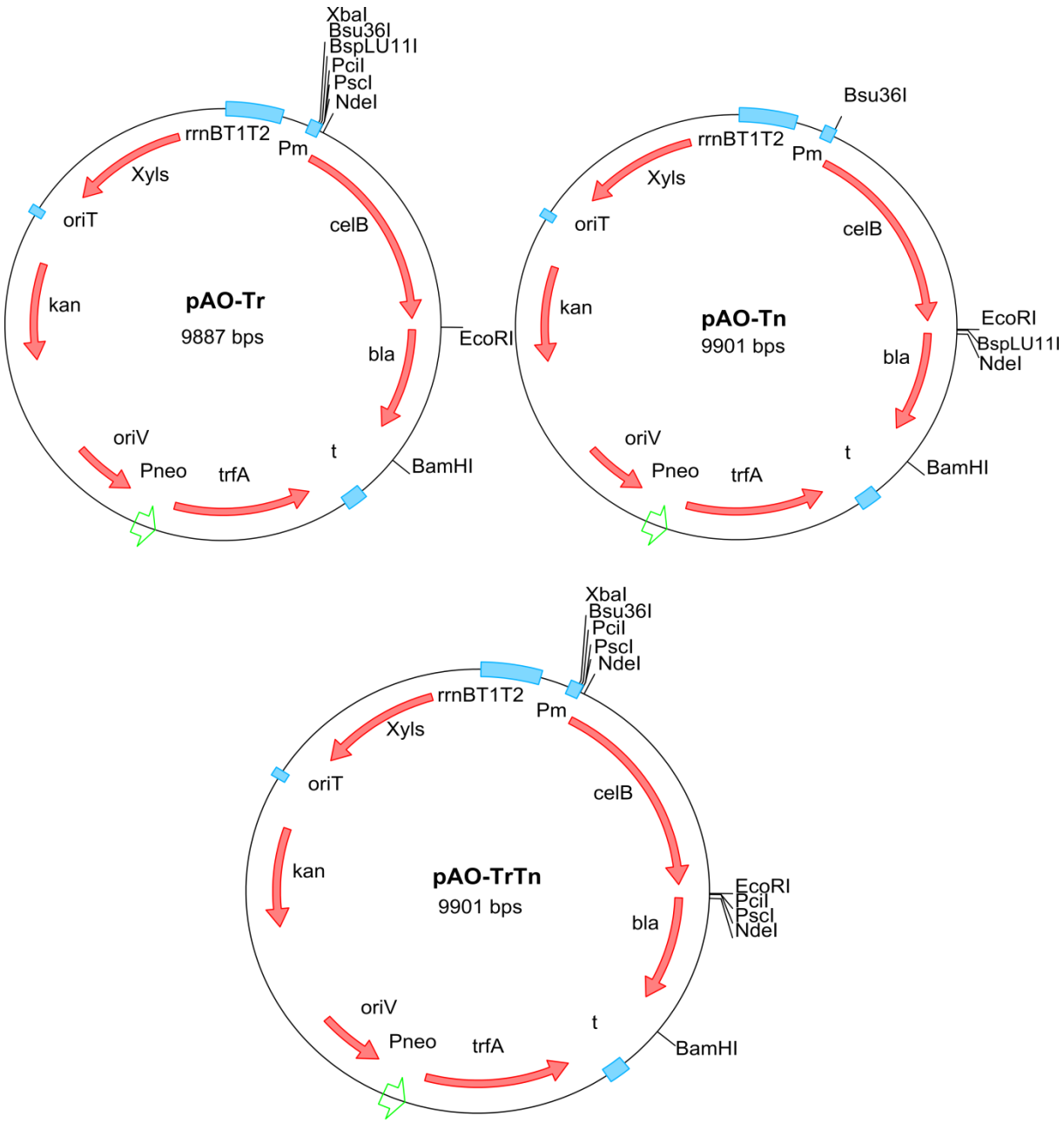

Figure 2-2: Plasmid map of pAO-Tr, pAO-Tn and pAO-TrTn. All containing the *xylS/Pm* expression cassette and *celB* and *bla* are co-transcribed as reporters for Pm. Kanamycin is used as selective antibiotic during culturing.

2.1.3 Oligonucleotides used for plasmid construction and as primers

The synthetic oligonucleotide sequences presented in table 2.2 were used to create plasmid constructs with alternative 5'-UTRs to the wild type 5'-UTR. The sequences were obtained from previously conducted screenings of 5'-UTR libraries (See section 1.2) with the exception of the Long UTR max sequences which are obtained from the bioinformatics tool RBS calculator (section 2.2.4). The oligonucleotides were designed using Clone Manager Suite (Sci-Ed Software). The forward and reverse strands were annealed as described in section 2.2.2.

Table 2-2: An overview of the synthetic oligonucleotides used in this study to construct plasmids with alternate 5'-UTRs, containing oligonucleotide name, sequence in 5'-end to 3'-end direction, length, the restriction sites overhangs designed at the 5'-end and 3'-end (listed in that order) and the constructs made by the specific oligonucleotide. The pLUTR constructs are denoted with the upstream region of the 5'-UTR fusion first followed by the downstream region.

Name	**Sequence (5' to 3')**	**bp**	**Restriction sites**	**Used to construct**
LUTR wt fwd	CATGTACAATAATAATGGAGTCATGAACATA TCATACACTAGTATATACTAGATATGTACAA TAATAATGGAGTCATGAACA	82	PciI, NdeI	pLUTR
LUTR wt rev	TATGTTCATGACTCCATTATTATTGTACATAT CTAGTATATACTAGTGTATGATATGTTCATGA CTCCATTATTATTGTA	80	PciI, NdeI	
LUTR LV-1 fwd	CATGTACCATTATAACGGAGTAATGAACATA TCATACA	38	PciI, SpeI	pLUTR LV-1 wt
LUTR LV-1 rev	CTAGTGTATGATATGTTCATTACTCCGTTATA ATGGTA	38	PciI, SpeI	pLUTR LV-1 Tn24
LUTR LV-2 fwd	CATGTACcATAATACAGGAGTTATGAACATA TCATACA	38	PciI, SpeI	pLUTR LV-2 wt
LUTR LV-2 rev	CTAGTGTATGATATGTTCATAACTCCtGTATT ATGGTA	38	PciI, SpeI	pLUTR LV-2 Tn25

LUTR tr28 fwd	CATGTATATAATAATGGAGAAATGAACATATCATACA	38	PciI, SpeI	pLUTR Tr28 wt
LUTR tr28 rev	CTAGTGTATGATATGTTCATTTCTCCATTATTATATA	38	PciI, SpeI	pLUTR Tr28 Tn47
LUTR tn24 fwd	CTAGTATATACTAGATATGTACATTAATATAGGAGTCATCAACA	44	SpeI, NdeI	pLUTR wt Tn24
LUTR tn24 rev	TATGTTGATGACTCCTATATTAATGTACATATCTAGTATATA	42	SpeI, NdeI	pLUTR LV-1 Tn24
LUTR tn25 fwd	CTAGTATATACTAGATATGTACAATGATAATGGAGTCATAAACA	44	SpeI, NdeI	pLUTR wt Tn25
LUTR tn25 rev	TATGTTTATGACTCCATTATCATTGTACATATCTAGTATATA	42	SpeI, NdeI	pLUTR LV-2 Tn25
LUTR tn47 fwd	CTAGTATATACTAGATATGTAATAAACTAAAGGAGTTATGAACA	44	SpeI, NdeI	pLUTR wt Tn47
LUTR tn47 rev	TATGTTCATAACTCCTTTAGTTTATTACATATCTAGTATATA	42	SpeI, NdeI	pLUTR Tr28 Tn47
LUTR tn58 fwd	CTAGTATATACTAGATATGTACATTAACAAAGGAGTCATATACA	44	SpeI, NdeI	pLUTR wt Tn58
LUTR tn58 rev	TATGTATATGACTCCTTTGTTAATGTACATATCTAGTATATA	42	SpeI, NdeI	pLUTR LV-2 Tn58
LUTR LII-11 fwd	CTAGTATATACTAGATATGTTACACAATAATGGAGTAATGAACA	44	SpeI, NdeI	pLUTR wt LII-11
LUTR LII-11 rev	TATGTTCATTACTCCATTATTGTGTAACATATCTAGTATATA	42	SpeI, NdeI	pLUTR LV-2 LII-11
LUTR DI-3 fwd	CTAGTATATACTAGATATGTGGCATAATAATGGAGTTATGCACA	44	SpeI, NdeI	pLUTR wt DI-3
LUTR DI-3 rev	TATGTGCATAACTCCATTATTATGCCACATATCTAGTATATA	42	SpeI, NdeI	pLUTR LV-2 DI-3

LUTR DI-8 fwd	CTAGTATATACTAGATATGTCCCATAATAATGGAGTCATGAACA	44	SpeI, NdeI	pLUTR wt DI-8
LUTR DI-8 rev	TATGTTCATGACTCCATTATTATGGGACATATCTAGTATATA	42	SpeI, NdeI	pLUTR LV-2 DI-8
LUTR max 1 fwd	CTAGTAGATCTAATACACCCAATCTTTTTAGAGGAGGTTTTACA	44	SpeI, NdeI	pLUTR LV-2 max 1
LUTR max 1 rev	TATGTAAAACCTCCTCTAAAAAGATTGGGTGTATTAGATCTA	42	SpeI, NdeI	
LUTR max 2 fwd	CTAGTATACACGTTCAGCAAGAGCAACCGCCGAGGAGGTTAACA	44	SpeI, NdeI	pLUTR LV-2 max 2
LUTR max 2 rev	TATGTTAACCTCCTCGGCGGTTGCTCTTGCTGAACGTGTATA	42	SpeI, NdeI	
LUTR max 3 fwd	CTAGTAGGCGCAACCACCAGTTCAGCAAATCAGGAGGTTCTCCA	44	SpeI, NdeI	pLUTR LV-2 max 3
LUTR max 3 rev	TATGGAGAACCTCCTGATTTGCTGAACTGGTGGTTGCGCCTA	42	SpeI, NdeI	
LUTR max 4 fwd	CTAGTAATCCCAGTTTACGCCGACCATACCCTAAGGAGGTCCCA	44	SpeI, NdeI	pLUTR LV-2 max 4
LUTR max 4 rev	TATGGGACCTCCTTAGGGTATGGTCGGCGTAAACTGGGATTA	42	SpeI, NdeI	
LUTR max 5 fwd	CTAGTAGCAACAGGTCCTATACAACTAACCTAAGGCAGGTATCA	44	SpeI, NdeI	pLUTR wt max 5
LUTR max 5 rev	TATGATACCTGCCTTAGGTTAGTTGTATAGGACCTGTTGCTA	42	SpeI, NdeI	
LUTR max 6 fwd	CTAGTAGCCTTTCAGCCTCAGCCCAGAACCTTTAAGGAGGTACA	44	SpeI, NdeI	pLUTR wt max 6
LUTR max 6 rev	TATGTACCTCCTTAAAGGTTCTGGGCTGAGGCTGAAAGGCTA	42	SpeI, NdeI	
LUTR max 7 fwd	CTAGTAACACGTCCCCCAATAGTTATTTCTTAAGGAGGTCCCCA	44	SpeI, NdeI	pLUTR wt max 7
LUTR max 7 rev	TATGGGGACCTCCTTAAGAAATAACTATTGGGGGACGTGTTA	42	SpeI, NdeI	
LUTR max 8 fwd	CTAGTACCGCCACTAATAGTCCGCGCCCTTAAGGAGGAATCCCA	44	SpeI, NdeI	pLUTR wt max 8
LUTR max 8 rev	TATGGGATTCCTCCTTAAGGGCGCGGACTATTAGTGGCGGTA	42	SpeI, NdeI	

The primer sequences utilized for sequencing and qRT-PCR (section 2.2.3) are listed in table 2-3

Table 2-3: List of the primers used for sequencing and in qRT-PCR transcript analysis

Name	**Sequence (5' to 3')**	**Target**
Sequencing		
552F	AACGGCCTGCTCCATGACAA	*bla* 5'-UTR region
UTR.bla.rev	CAAGGATCTTACCGCTGTTG	*bla* 5'-UTR region
qRT-PCR		
Bla fwd	ACGTTTTCCAATGATGAGCACTT	*bla*
Bla rev	TGCCCGGCGTCAACAC	*bla*
16S fwd	ATTGACGTTACCCGCAGAAGAA	16S rRNA
16S rev	GCTTGCACCCTCCGTATTACC	16S rRNA

2.1.4 Enzymes

All enzymes and buffers for enzymatic reactions were ordered from New England Biolabs®*inc*.

2.2 Methods

2.2.1 Cultivation and manipulation of bacteria

E. coli DH5α was cultivated in liquid LB or on solid LA plates. Selective antibiotics were added if required.

Competent cells

Competent cells are susceptible for uptake of plasmid DNA. This can be achieved by exposing the cells to certain cations, such as Mn^{2+}, Ca^{2+} and Rb^{+} followed by heat shock treatment (Aune et al. 2010).

E. coli DH5α competent cells were first grown in 10 mL Psi medium in an Erlenmeyer flask at 37 ºC, 225 rpm ON. The following day 2 mL of the ON culture were transferred to 200 mL Psi medium in a 500 mL Erlenmeyer flask and grown to OD_{600} = 0.4 at 37 ºC 225 rpm. Then the cell culture was incubated on ice for 15 minutes, followed by centrifugation (4 ºC, 4500 rpm, 5 minutes). The supernatant was removed and the pellet was carefully dissolved in 80 mL TFB1 followed by another round of centrifugation. The supernatant was removed and the pellet was carefully dissolved in 6 mL TBF2. The cell solution was then aliquoted in volumes of 100 and 500 µL and snap-frozen using ethanol cooled with dry ice. The aliquots were stored at -80 ºC.

Transformation

Competent cells were thawed on ice, then 10 µL plasmid/ligation mixture was added to 100 µL cell solution and left on ice for 20-30 minutes. The cells were exposed to heat shock by placing them in a 42 ºC water bath for 42 seconds and then put back on ice for 2 minutes.

1 mL LB was added followed by incubation at 37 ºC (225 rpm) for 1 to 1.5 hours. 100 µL cell solution (diluted or concentrated if needed) spread on LA plates containing selective antibiotics and incubated at 37 ºC ON.

Glycerol stock

For long term storage of cell cultures 3 mL LB + selective antibiotic were inoculated with a single colony and incubated at 37 ºC, 225 rpm ON. 1 mL culture was mixed with 0.5 mL 60% glycerol and stored at -80 ºC.

Plasmid isolation

Plasmids were isolated using the Wizard® *Plus SV* Miniprep DNA Purification System from Promega (the centrifugation part of the protocol). Single colonies were transferred to 13 mL tubes with 3 mL LB containing selective antibiotics and incubated at 30 ºC, 225 rpm for 16 – 18 hours.

- The next morning cells were centrifuged for 5 minutes at 10 000 x g.
- The pellet was resuspended in 250 µL Cell Resuspension Solution by vortexing and then transferred to a sterile 1.5 mL Eppendorf tube.
- 250 µL Cell Lysis Solution was added and mixed by inverting the tube 4 times followed by incubation for 5 minutes.
- 10 µL of Alkaline Protease Solution was added and mixed by inverting the tube 4 times followed by incubation for 5 minutes.
- 350 µL of Neutralization Solution was added and mixed by inverting the tube 4 times.
- The lysate was centrifuged for 10 minutes at 13 000 rpm.
- The cleared lysate was transferred to a spin column placed in a collection tube followed by centrifugation for 1 minute at 13 000 rpm.
- The flow-through was discarded and 750 µL Column Wash Solution was added followed by centrifugation for 1 minute at 13 000 rpm.
- The flow-through was discarded and the wash procedure was repeated using 250 µL column wash solution followed by centrifugation for 2 minutes at 13 000 rpm.
- The spin column was transferred to a new sterile Eppendorf tube and the plasmid was eluted by adding 100 µL Nuclease-free water followed by centrifugation for 1 minute at 13 000 rpm.
- Plasmid yield was determined by spectrophotometry (NanoDrop 1000, Thermo Scientific) and stored at -20 ºC.

For high plasmid yields the NucleoBond® Xtra Midi/Maxi (Macherey-Nagel) was used (Midi + low-copy protocol).

- A start culture was made by inoculating 3 mL LB containing selective antibiotics with single colonies followed by incubation for 8 hours at 37 ºC.
- An ON culture was made by diluting the start culture 1/100; two times 750 µL start culture in 75 mL LB + antibiotics in 500 mL Erlenmeyer flask making 150 mL total volume. The ON culture was incubated at 37 ºC, 225 rpm.

- The following day the two 75 mL cultures were combined and centrifuged at 5000 x g for 15 minutes at 4 ºC.
- The pellet was resuspended in 16 mL Resuspension Buffer RES + RNase A by vortexing.
- 16 mL Lysis Buffer LYS was added and mixed by inverting the tube 5 times.
- The column filter was prepared by placing it in a NucleoBond®Xtra Column and equilibrating by adding 12 mL Equilibration Buffer EQU as instructed in the protocol.
- 16 mL Neutralization Buffer NEU was added to the lysate and mixed by inverting the tube 15 times.
- Following the "alternative" part of step 8 in the protocol, the precipitate was removed by centrifugation at 5000 x g for 15 minutes to prevent clogging of the filter.
- The lysate was applied to the equilibrated NuceoBond®Xtra Column Filter.
- The NuceoBond®Xtra Column and Column Filter were washed by adding 5mL Equilibration Buffer EQU.
- The column filter was discarded and the column was washed with 8 mL Wash Buffer WASH.
- The plasmid DNA was eluted by adding 5 mL Elution Buffer ELU. The eluate was collected in a 50 mL centrifuge tube.
- 3.5 mL isopropanol (room temperature) was added to precipitate the eluted plasmid DNA followed by vortexing and then left on the bench for 2 minutes. The mixture was then centrifuged at 10 000 x g for 30 minutes at 4 ºC. The supernatant was removed carefully.
- 2 mL 70% ethanol was added to the pellet followed by centrifugation at 10 000 x g for 15 minutes at room temperature. The ethanol was carefully removed using a vacuum pump and the pellet was left to dry at room temperature for 5-10 minutes.
- The pellet was dissolved in 200 µL distilled water. Plasmid yield was determined by spectrophotometry (NanoDrop 1000, Thermo Scientific) and the plasmids were stored at -20 ºC.

Culturing and sampling for β-lactamse assays and transcript measurement

For each strain two 13 mL tubes containing 2 mL LB + selective antibiotics each, were inoculated with single colonies from a fresh agar plate and incubated at 30 ºC ON. The next

morning, the OD_{600} was measured and the parallels with the most even results were diluted to make a start culture of OD_{600} = 0.05 in 15 mL LB + selective antibiotics in a 125 mL Erlenmeyer flask. The cultures were grown at 37 ºC, 225 rpm until OD_{600} reached 0.1. The cultures were immediately induced with 2 mM *m*-toluic acid and incubated at 30 ºC, 225 rpm for 5 hours.

For RNA isolation 0.5 mL culture were mixed with 1 mL RNAprotect (QIAGEN) by vortexing for 5 seconds followed by 5 minutes incubation at room temperature. The samples were centrifuged for 10 minutes at 5 000 rpm, 4 ºC. The supernatant was discarded and the pellet was stored at -80 ºC. For β-lactamase assays 5 mL of each cell culture were transferred to a 13 mL tube followed by centrifugation for 10 minutes at 8000 rpm, 4 ºC. The supernatant was discarded and the pellet was resuspended in 5 mL assay buffer followed by an additional centrifugation step. The supernatant was discarded and the pellet was stored at -20 ºC.

2.2.2 DNA manipulation

DNA digestion by type II Restriction endonucleases

Type II restriction endonucleases recognize specific DNA sequences (often palindromic) of 4-8 bp and cleave the DNA within or close to the recognized sequence (Pingoud et al. 2001). Some endonucleases leave "sticky ends" or overhang of DNA bases making it possible for cleaved DNA fragments with matching overhang to be ligated together(Roberts 2005).

The reaction conditions used were obtained using the double digest finder tool (Biolabs) provided by New England Biolabs® inc.

Reaction mixture e.g 30 µL:

- 7 µL DNA
- 3 µL 10 x recommended NEBuffer
- 3 µL 10 x BSA if recommended
- 1 µL Restriction enzyme A
- 1 µL Restriction enzyme B
- Distilled water to make a total of 30 µL

The reaction mixture was incubated in a water bath at 37 ºC for 2 hours. The resulting DNA fragments were separated using gel electrophoresis and required fragments were cut out of the gel and purified using the Qiagen gel extraction kit (described below).

Used for oligonucleotide insertion:

Reaction mixture e.g. 80 µL:

- 20 µL DNA
- 8 µL 10 x recommended NEBuffer
- 0.8 µL 100 x BSA if recommended
- 2.5 µL Restriction enzyme A
- 2.5 µL Restriction enzyme B
- Distilled water to make a total of 80µL

The reaction mixture was incubated in a water bath at 37 ºC ON followed by heat inactivation at the conditions given by the New England Biolabs®$_{inc}$ homepage for the relevant enzymes. 4 µL of CIP (Alkaline Phosphatase, Calf Intestinal) was added and the sample was incubated in a water bath at 37 ºC for 2 hours. The digested product was then purified using the Qiagen PCR purification kit protocol (using the microcentrifuge alternative), and used as vectors for oligonucleotide insertion.

Gel extraction

After digested DNA fragments were separated using gel electrophoresis the required fragments were cut out of the agarose gel using a clean scalpel, placed in a sterile Eppendorf tube and purified using the QIAquick® Gel Extraction Kit.

- Buffer QG was added so that the gel was covered.
- The tube was incubated at 42 ºC in a water bath until the gel was dissolved. Vortexing was preformed if necessary.
- 100 µL isopropanol were added and mixed with the dissolved gel slice.
- A QIAquick spin column was placed in a 2 mL collection tube and the sample was applied to the column followed by 1 minute centrifugation at 13 000 rpm.

- The flow-through was discarded and 720 µL wash buffer PE was added to the column followed by 1 minute of centrifugation at 13 000 rpm.
- The flow-through was discarded and the sample was centrifuged one additional time for 2 minutes to remove any remaining wash buffer.
- The column was placed in a sterile Eppendrof tube and the DNA was eluted by applying 32 µL of distilled water to the column followed by centrifugation for 1 minute at 13 000 rpm.
- The DNA yield was determined by the use of a NanoDrop 1000 (Thermo Scientific) spectrophotometer.

Annealing of synthetic oligonucleotides

Synthetic oligonucleotides (tab. 2-2) were ordered from Sigma-Aldrich®. They are synthesized as separate single-stranded DNA strands, one forward strand and one reverse strand, and thus need to be annealed before use. The synthetic oligonucleotides do not have a 5' phosphate and is therefore treated with T4 polynucleotide kinase (PNK) which synthesizes the transfer of a phosphate from ATP to the 5'-hydroxyl termini of polynucleotides (Richardson 1965).

Reaction mixture:

- 7 µL forward oligonucleotide [100 µmol/µL]
- 7µL reverse oligonucleotide [100 µmol/µL]
- 2 µL Buffer T4 DNA ligase
- 0.7 µL PNK

The reaction mixture was incubated at 37 ºC for 30 minutes followed by heat inactivation by incubating the reaction mix at 65 ºC for 30 minutes. 4µL 1M NaCl was added to the reaction mixture and the oligonucleotides were annealed in an Eppendorf Mastercycler® PCR machine using the "ANNEAL" program (appendix A, table 8-1).

Ligation reaction

Ligase catalyzes the formation of phosphodiester bonds between directly adjacent 3'-hydroxyl and 5'-phosphoryl termini in DNA and is therefore used as a tool for attaching DNA fragments with matching overhangs (Lehman 1974).

To calculate the amount of vector and insert to add, the following equation (Eq. 1) was used with a ratio of 3.

$$Equation\ 1: ng\ insert = \frac{ng\ vector\ \times kb\ insert\ \times ratio}{kb\ vector}$$

A sample containing only the vector DNA served as religation control.

Ligation mixture e.g. 20 µL

- 1 µL T4 DNA ligase diluted 1/10 with diluent A (NEB)
- 2µL T4 DNA ligase buffer
- 17 µL Vector/insert

The ligation mixture was incubated at 4 ºC ON.

DNA sequencing

Isolated plasmids were transferred to sterile Eppendorf tubes in volume amounts containing 1000 ng plasmid DNA. The plasmids were dried on a heating block at 50 ºC until about 15 µL liquid were left. The tubes were then sent to Eurofins MWG Operon for sequencing using primers listed in table 2-3.

2.2.3 Expression analysis

Testing of ampicillin resistance by replica plating

To identify the level of ampicillin resistance acquired through recombinant expression of β-lactamase, colonies were grown on a series of large agar plates with an increasing ampicillin concentration.

A colony was transferred from a fresh agar plate to a 96-well plate (Nunc) containing 100 µL LB + selective antibiotics in each well. 4-6 parallels were used for each strain. The 96-well plate was sealed with parafilm and incubated on a horizontal shaker board at 30 ºC, 800 rpm for about 24 hours.

The next day, two sets of big agar plates (50 mL LA) with varying ampicillin concentration were made, one with and one without inducer. In addition a plate containing kanamycin was prepared as a positive control. Agar, antibiotic +/- inducer were added to a 50 mL centrifugation tube and mixed by inverting the tube 4 times before pouring the plates. The plates were left on the laboratory bench until the agar was solid and then dried in a clean bench.

When the agar plates were dry the cell cultures were diluted by transferring the cultures to a new 96-well plate containing 100 µL LB medium in each well using a 96-pin replicator (transferring about 1 µL). This was repeated one additional time resulting in a dilution of about 1: 10 000. The diluted cultures were transferred to the agar plates using the 96-pin replicator. The plates were incubated at 30 ºC for two days.

β-lactamase assay

β-lactamase breaks the β-lactam ring of β-lactam antibiotics such as penicillin. This changes the UV absorption spectrum making it possible to use a spectrophotometer to measure the enzymatic activity of β-lactamase (Waley 1974).

The frozen pellet was resuspended in 2 mL assay buffer. The cells were lysed by sonication using the tapered microtip of a Branson Sonifier 250 for 2.5 minutes at 30% duty cycle, 3.0 output control. The sample was kept on ice during sonication and then centrifuged for 10 minutes at 8000 rpm, 4 ºC. The supernatant constituted the sample.

For measuring the reaction kinetics of β-lactamase in the samples several dilutions were made and applied to a 96-well UV Transparent Nunc plate in three parallels with the following reaction mixture:

- 20 µL of sample
- 170 µL assay buffer
- 10 µL penicillin-G (pen-G) stock (10 mg/mL)

In addition a standard was made consisting of a blank sample and 100 through 500 µg/mL with intervals of 100 µg/mL pen-G.

Samples were measured immediately after pen-G was added using the following settings: 15 seconds mixing before first read, followed by absorbance measurements at 240 nm every 30 seconds for 10 minutes.

Determination of the protein concentration:

Coomassie Brilliant Blue G-250 binds proteins and by doing so shifts its absorption maximum from 465 to 595 nm (Bradford 1976). Absorbance and protein amount follow a linear relationship between the range 0.05 and 0.5 mg/mL using the microtiter plate protocol (Bio-Rad).

The total protein amount was determined using the Bio-Rad protein assay procedure (standard microtiter plate protocol). First the dye reagent was diluted by mixing 1 part dye reagent concentrate with 4 parts distilled water followed by filtration through a Whatman # 1 filter. 6 dilutions of the protein standard were prepared (0.05, 0.1, 0.2, 0.3, 0.4, and 0.5 mg/mL BSA). 10 µL of each sample and protein standard in addition to a blank sample were added to separate wells of the microtiter plate (3 parallels of each sample and standard). Diluted dye reagent (200 µL) was added to each well followed by mixing and incubation for 5 minutes prior to absorbance measurement at 595 nm.

Transcript analysis

Cultures were grown, as described in section 2.2.1, 5 hours post induction with 2 mM *m*-toluic acid.

RNA isolation and cDNA synthesis

Total RNA was isolated using the RNAqueous® Kit (Ambion). Samples were kept on ice. All centrifugations were performed at 13 000 rpm for 1 minute.

- The cell pellet was resuspended in 100 µL TE-buffer. A small amount of lysozyme was added using a pipette tip followed by mixing and incubation for 5 minutes at room temperature.
- 300 µL Lysis/Binding Solution was added and mixed by vortexing.

- 400 µL 64% ethanol was added and mixed by inverting the tube 4 times.
- The lysed sample was transferred to the filter column placed in a collection tube and centrifuged.
- The flow-through was discarded and 700 µL Wash Solution 1 was added followed by centrifugation.
- The flow-through was discarded and 500 µL Wash Solution 2/3 was added followed by centrifugation. This step was performed twice.
- The flow-through was discarded followed by an additional centrifugation step.
- The sample was eluted by adding 50 µL preheated (80 ºC) Elution Solution followed by centrifugation. The step was repeated using 25 µL Elution Solution.
- 2 µL of the sample was transferred to a new Eppendorf tube for measuring of RNA concentration using the NanoDrop 1000 (Thermo Scientific).

DNA was removed from the sample using the TURBO DNA-*free*™ Kit (Ambion).

- RNA (3000 ng) was mixed with RNase free water making a volume of 25 µL.
- 2.7 µL TURBO DNA-free buffer and 1 µL Ambion Turbo DNase was added followed by incubation for 30 minutes at 37 ºC.
- 5 µL Inactivation Mixture was added followed by incubation for 2 minutes at room temperature.
- The sample was centrifuged at 13 000 rpm for 2 minutes. 2 µL were transferred from the top of the sample to 18 µL RNase free water.

The cDNA was synthesized using the First-Strand cDNA Synthesis Kit (Ambion).

- The sample was heated for 10 minutes at 65 ºC.
- 3.5 µL bulk mixture (2.5 µL Bulk reaction mix, 0.5 µL hexamere primers, 0.5 µL DTT) were mixed with 4 µL sample.
- The samples were incubated for 1 hour at 37 ºC.

Quantitative Real-Time Polymerase Chain Reaction (qRT-PCR)

SYBR green binds double stranded DNA and the resulting complex emits green florescent light when exited. This property is utilised for monitoring dsDNA amount in real time during polymerase chain reaction (PCR)(Zipper et al. 2004).

The following components were mixed in each well of qRT-PCR strips:

- 12.5 µL Power SYBR® Green master mix
- 2.5 µL Forward primer (2 mg/mL)
- 2.5 µL Reverse primer (2 mg/mL)
- 2.5 µL Distilled water
- 5 µL Sample

3 parallels were made for each sample in addition to 3 non template controls. Primer sequences for *bla* and 16S rRNA (endogenous control) are shown in table 2-3. Samples for the endogenous control were diluted 1:400 and target samples were diluted 1:100. The qRT-PCR strips containing reaction mixture were spun down before running qRT-PCR using an Applied Biosystems 7500 Real Time PCR system with the standard settings.

2.2.4 Bioinformatics analysis

Clone Manager professional suite (Sci Ed software)

The Clone Manger suite used (containing Clone manager 6, Align plus 4 and Primer Designer 4) allows the managing of DNA sequences *in silico*, including enzymatic digestion, ligation and alignment of two or more sequences. The program was used for planning of cloning steps, design of synthetic oligonucleotides, making sequence alignments, sequence analysis and creation of plasmid maps.

RBS calculator$_{v1.1}$ (Salis lab, Penn state university)

The Ribosome Binding Site (RBS) Calculator is a bioinformatics tool engineered for the prediction of the translation initiation rate (TIR) of a RBS (reverse engineering) or the design

of a RBS sequence with the desired TIR (forward engineering) in bacteria (Salis 2011). The software utilizes an equilibrium statistical thermodynamic model which quantifies the strengths of molecular interaction between mRNA transcript and the 30S ribosome construct to predict the TIR (Salis et al. 2009; Salis 2011).

The software was used to evaluate TIRs for 5'-UTR variants obtained through screening and the 5'-UTR fusions involving these, by using the reverse engineering option with the 5'-UTR plus the 50 first coding base pairs of *bla* (appendix B) as mRNA input and ACCTCCTTA as 16S rRNA input. 8 sequences were designed (LUTR max 1-8 table 3-2) using the forward engineering with constraints option. The following sequence was used as RBS constraints input:

ATACACTAGTANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNCAT

where N denotes the nucleotides to be chosen by the software. The 50 first coding nucleotides of *bla* (appendix B) were used as input for protein coding sequence and ACCTCCTTA were used as 16S rRNA input. For maximum TIR the option "Goal: Maximize" was chosen.

Quickfold

Quickfold is a web based tool and part of the Mfold web server, which consists of different software made for prediction of the secondary structure of single stranded nucleic acids. The softwares are based on a minimum free energy algorithm for folding of the input sequence (Zuker 2003). The software allows for prediction of several sequences at once and returns a table of the lowest free energies in kcal/mol for each structure as well as a pdf-file with the structure models. Default settings were used except for the RNA 3.0 energy rule option and maximum distance between paired bases of 30 to only allow local structures.

3 Results

3.1 Characterization of the transcription stimulating and translation stimulating traits of the short 5'-UTR variants obtained by selective screening

3.1.1 Exploring the transcription stimulating and translation stimulating qualities by switching position of the two functional 5'-UTR units in the pAO double construct

Several transcription stimulating 5'-UTR variants (Tr) and translation stimulating 5'-UTR variants (Tn) had prior to this study been identified (tab. 1-1) by screening for increased *bla* expression by altering the 5'-UTR in front of *celB* and *bla*, respectively (section 1.2). However, the effect of the Tr 5'-UTR variants on translation and the Tn 5'-UTR variants on transcription was not known. In order to explore this, pAO constructs were made with switched 5'-UTR positions, placing the Tr 5'-UTR variants in the *bla* position and the Tn 5'-UTR variants in the *celB* position. If the Tr 5'-UTR variants only enhance the transcription one would expect no increase in *bla* expression when placed in front of *bla*. The same goes for the Tn 5'-UTR variants, if they only stimulate translation no increase in *bla* expression should be observed when placed in front of *celB*. The resulting *bla* expression was tested by replica plating. Strains harbouring the plasmid with the stimulating 5'-UTR variants in their original position were included as controls (fig. 3-1 and 3-2).

From the results presented in figure 3-1 it can be observed that the translation stimulating 5'-UTRs led to an increase in ampicillin resistance of about 8-20-fold in the *bla* position and unchanged or a weak decrease in ampicillin resistance in the *celB* position compared to the wild type 5'-UTR sequence. This indicates that the Tn 5'-UTR enhance the translation, but has no positive effect on the amount of mRNA transcript.

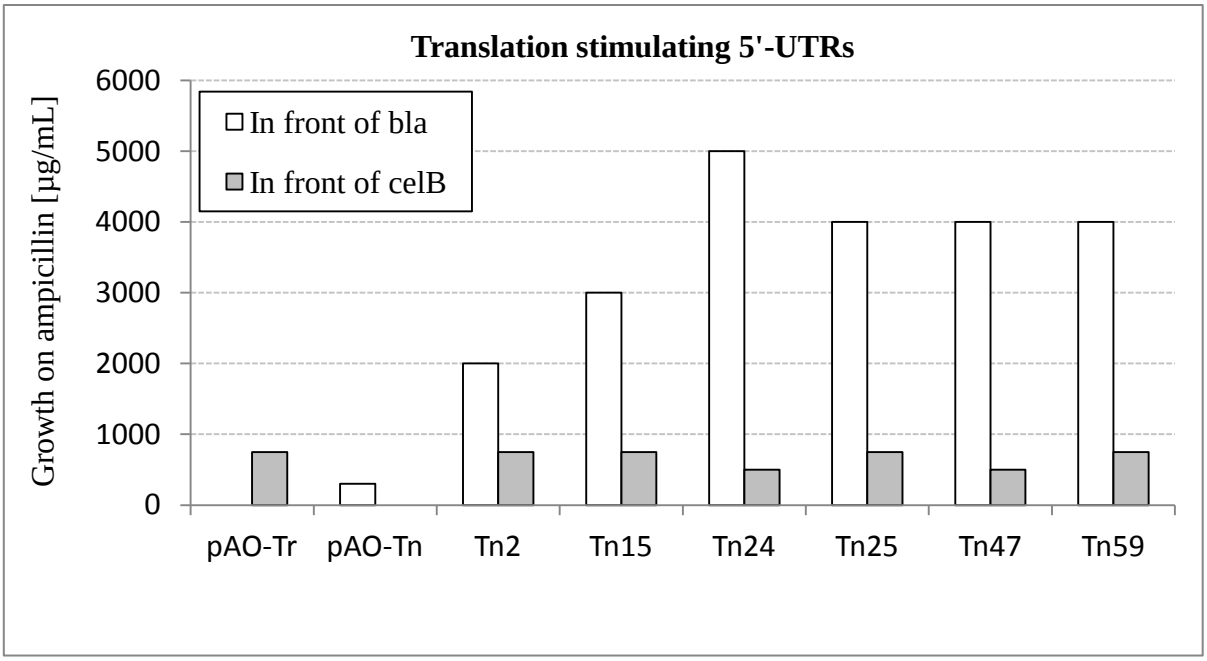

Figure 3-1: Replica plating of strains with the translation stimulating 5'-UTR variants placed either in front of *bla* (white columns) or in front of *celB* (grey columns) in the pAO double construct (fig 2-2). The two first columns show the resistance level of the strains harbouring the plasmids in figure 2-2 without any alteration of the 5'-UTR sequences. The 5'-UTR in the *bla* position will influence the *bla* expression at the translation level and the 5'-UTR in the *celB* position will influence the *bla* expression at the transcription level (fig. 1-3). The expression was induced with 0.5 mM *m* toluic acid. For the ampicillin concentrations of the plates used in this experiment see appendix C, table 8-2.

From the results of the replica plating presented in figure 3-2 it can be observed a ~ 2.5 fold increase in expression of the strains in which the Tr36 5'-UTR variant is placed upstream of *celB* and a weak increasing effect from the remaining Tr 5'-UTR variants in the *celB* position compared to the wild type. In the *bla* position the transcriptional stimulating 5'-UTRs also resulted in a stimulating effect, but on average lower than what was observed for the Tn 5'-UTR variants (fig 3-1). The results indicate that the 5'-UTR variants obtained by screening for enhanced transcription also have a positive effect on the translation.

It should be noted that the sequences of the pAO constructs with switched positions were not verified by sequencing.

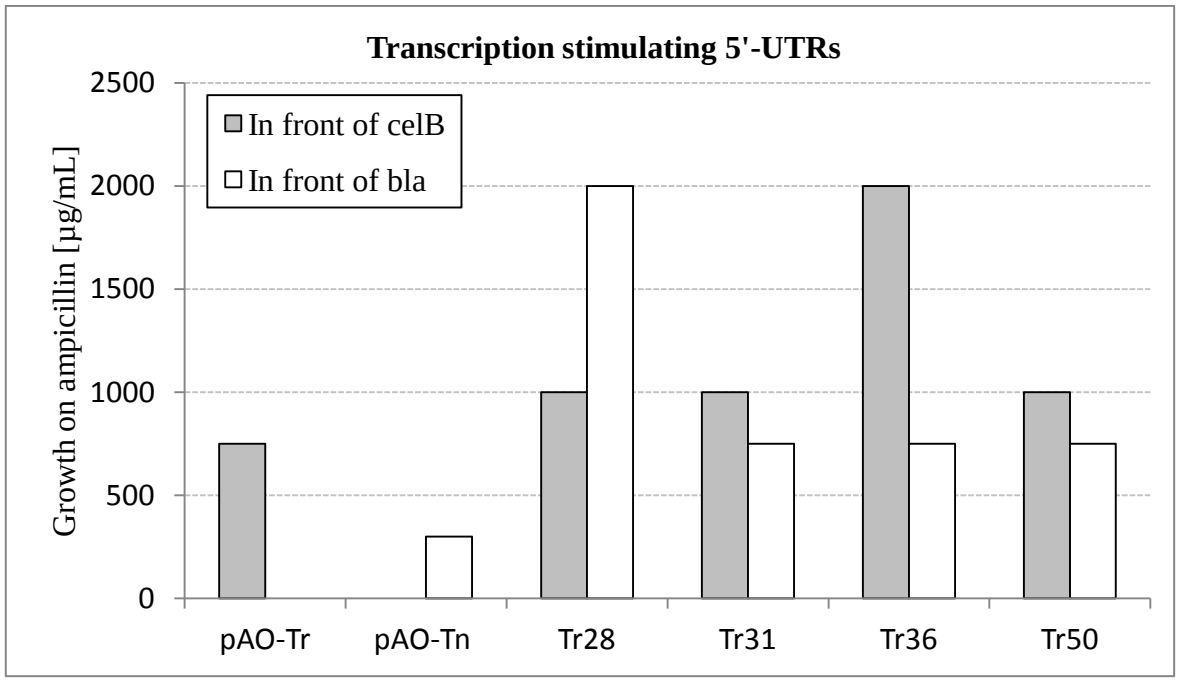

Figure 3-2: Replica plating of strains with the translation stimulating 5'-UTR variants placed either in front of *celB* (grey columns) or in front of *bla* (white columns) in the pAO double construct (fig 2-2). The two first columns show the resistance of the strains harbouring the plasmids in figure 2-2 without any alteration of the 5'-UTR sequences. The 5'-UTR in the *bla* position will influence the bla expression at the translation level and the 5'-UTR in the *celB* position will influence the *bla* expression at the transcription level. The expression was induced with 0.5 mM *m* toluic acid. For the ampicillin concentrations of the plates used in this experiment see appendix C, table 8-2.

3.1.2 Exploring the transcription stimulating and translation stimulating qualities by β-lactamase assay and transcript analysis

Another method of evaluating the transcription and translation stimulating traits resulting from a 5'-UTR variant are by the use of β-lactamase assay and transcript analysis by qRT-PCR described in section 2.2.3. The measured β-lactamase activity adjusted to the total protein amount in a given cell culture gives information on the amount of translated β-lactamase in the strain. The transcript analysis gives information on the amount of *bla* transcript produced by a given strain and henceforth on the transcription efficiency compared to the other strains. It should be noted that the transcript amount will also be influenced by the mRNA decay rate in addition to the transcription efficiency. There might have been variance in mRNA decay rates between the different strains, but this was not explored in this study. β-lactamase assay and transcript analysis were used on strains harbouring some of the

transcription and translation stimulating 5'-UTR variants in front of *bla* in the pIB11 construct (fig 2-1). The LV-1 5'-UTR variant was included as a control since the same procedure had been utilized to characterize it in the published work of Berg et al (Berg et al. 2009). The results are presented in figure 3-3. The raw data from the assays and qRT-PCR as well as the protein standard curve used for determining protein concentration of the samples can be found in appendix D, table 8-5 to 8-9 and figure 8-1.

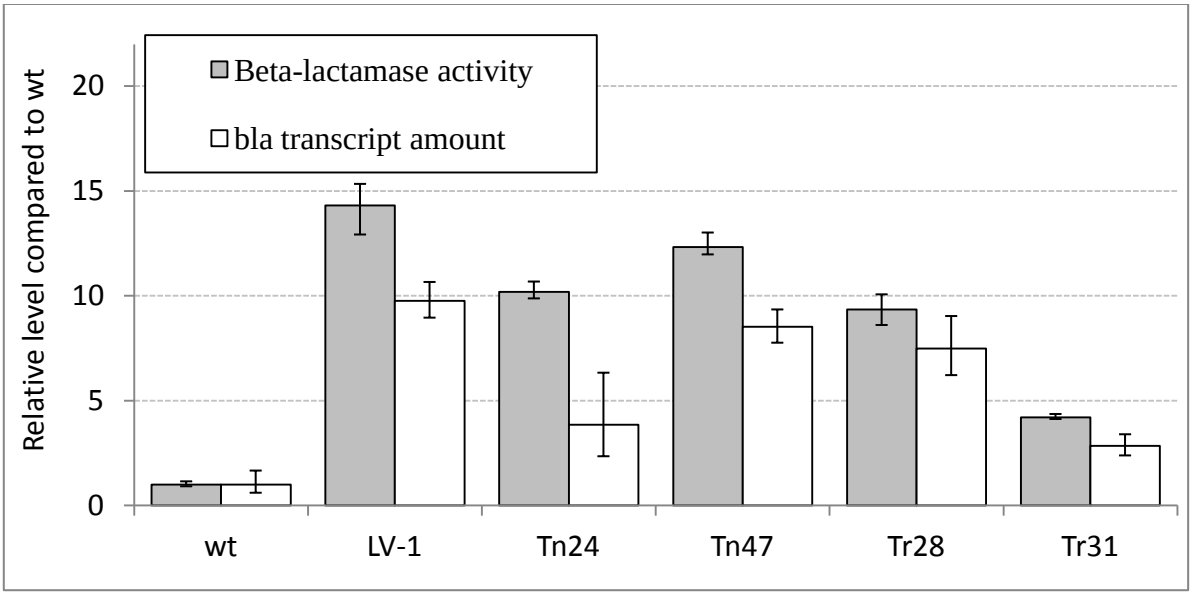

Figure 3-3: The β-lactamase activity (grey columns) and *bla* transcript amount (white columns) relative to the wt levels. The methods are described in section 2.2.3. The raw data used can be found in appendix D, table 8-5 to 8-9 and figure 8-1.

The LV-1 strain has previously been shown to increase both β-lactamase activity and the *bla* transcript amount by a factor of ~12 (Berg et al. 2009). The results for the LV-1 variant presented in figure 3-3 have some deviation from the previously obtained results with about 14-fold increase in β-lactamase activity and about 10-fold increase in *bla* transcript amount. The transcription stimulating Tr28 and Tr31 5'-UTR variants resulted in an increase in *bla* transcript of about 7-and 3-fold, respectively. However, the translation stimulating Tn24 and Tn47 5'-UTR variants also showed an increase in transcript amount of about 4- and 8-fold, respectively. The transcription stimulating effect of the Tn24 and Tn47 variant was not observed in the experiment using the pAO double construct (fig. 3-1).

Aside from the LV-1 variant, the Tn24 and Tn47 variants led to the highest β-lactamase activity with an increase of about 10- and 12-fold, respectively, followed by the transcription

variants Tr28 and Tr31 which led to an increase in β-lactamase activity of 9- and 4-fold, respectively. The Tr 5'-UTR variants showed an enhancing effect on both transcription and translation, but the effect on translation was lower than what was observed for the Tn 5'-UTR variants. This matches the results from the experiment using the pAO double construct (fig. 3-2).

3.2 Enhancement of *bla* expression by the use of 5'-UTR variants stimulating transcription and translation both as individual sequence elements and as combined 5'-UTR fusions

3.2.1 An additive effect on the *bla* expression is observed when enhancing both the transcription and the translation in the pAO-TrTn double construct

To test the possible additive effect on *bla* expression by stimulating both transcription and translation through 5'-UTR alterations, several pAO-TrTn (fig. 2-2) constructs were made. The constructs consisted of both transcription stimulating 5'-UTRs (Tr and LV) in front of *celB* and translation stimulating 5'-UTRS (Tn) in front of *bla.* In addition control constructs were made containing either a transcriptional or translational stimulating 5'-UTR accompanied by the wild type 5'-UTR in the other position. The construct names are denoted with the name of the 5'-UTR variant in the *celB* position followed by the name of the 5'-UTR variant in the *bla* position (e.g. Tr28 Tn47 means that the Tr28 5'-UTR variant is in front of *celB* and the Tn47 5'-UTR variant is in front of *bla).* The resulting gene expression of *bla* was tested by replica plating as described in section 2.2.3. The results are presented in figure 3-4.

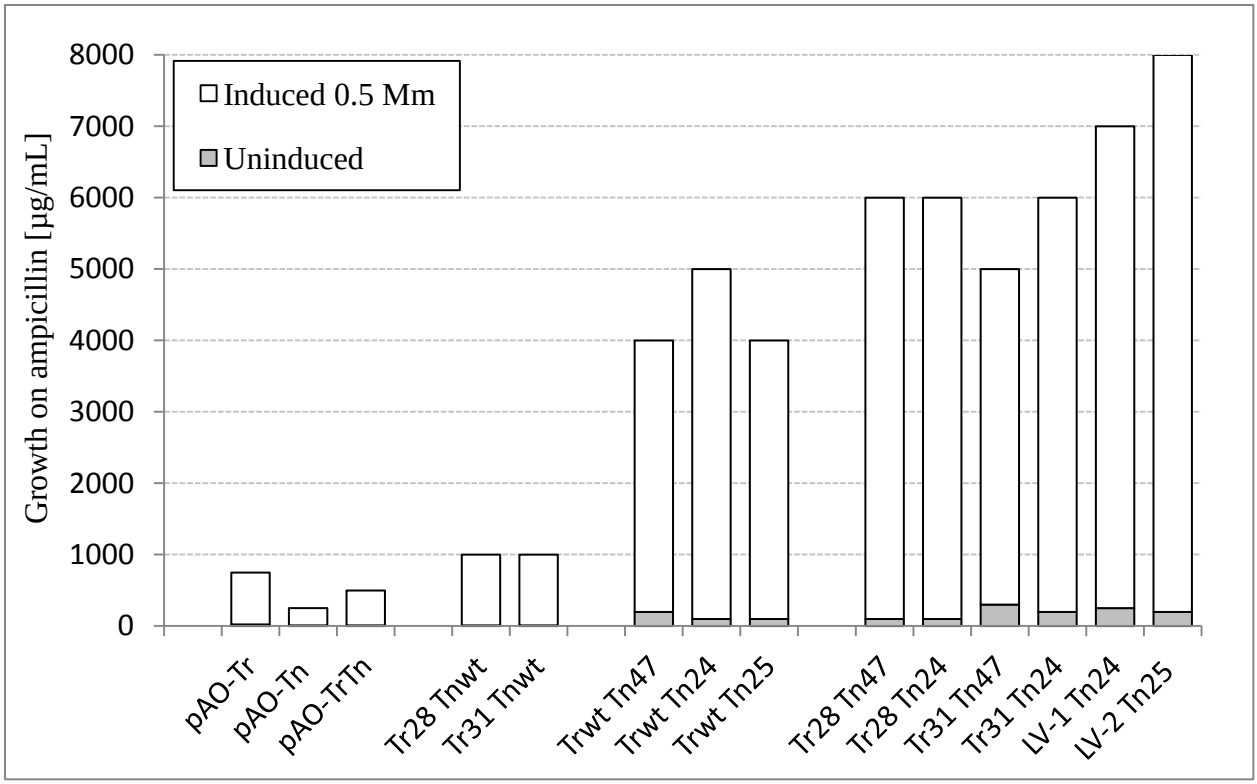

Figure 3-4: The bars show the maximum ampicillin tolerance level of the different strains during replica plating when inducing with 0.5 mM *m*-toluic acid. Growth at uninduced levels is shown in grey. The first four bars are from strains hosting the different plasmids shown in figure 2-2 without any alteration of the 5'-UTR sequences, followed by two strains with transcription stimulating variants in the *celB* position, three strains with translation stimulating variants in the *bla* position, and six strains with both transcription and translation stimulating 5'-UTRs in front of *celB* and *bla*, respectively. The maximum ampicillin concentration was 8000 µg/mL. For complete concentration range of the plates see appendix C, table 8-3.

The transcription stimulating 5'-UTRs (Tr) accompanied by the wt 5'-UTR in the *bla* position (no translation enhancement) only resulted in a weak increase in ampicillin resistance. However, by combining them with the more potent translation stimulating 5'-UTRs (Tn) in front of *bla* a level of resistance surpassing the ones achieved by stimulating just transcription or translation was observed (fig. 3-4). When combining the translation stimulating 5'-UTRs with the identified transcription stimulating 5'UTRs LV-1 and LV-2 from the published work of Berg and colleagues (Berg et al. 2009) an even larger increase in ampicillin resistance could be observed (fig. 3-4). The results indicate an additive effect on *bla* gene expression achieved by enhancing both the amount of *celB-bla* mRNA transcript and the *bla* translation efficiency.

These pAO-TrTn constructs were not confirmed by sequencing, but restriction analysis suggested a successful cloning procedure (fig. 3-5). When cut with the restriction enzyme NdeI correctly assembled pAO-TrTn construct would yield a fragment of 1775 bp and a fragment of 8126 bp, as opposed to the religated vector that would only yield one fragment of 9901 bp. Four parallels were tested for each of the TrTn combinations constructed and all of them had at least two parallels yielding the desired fragment pattern.

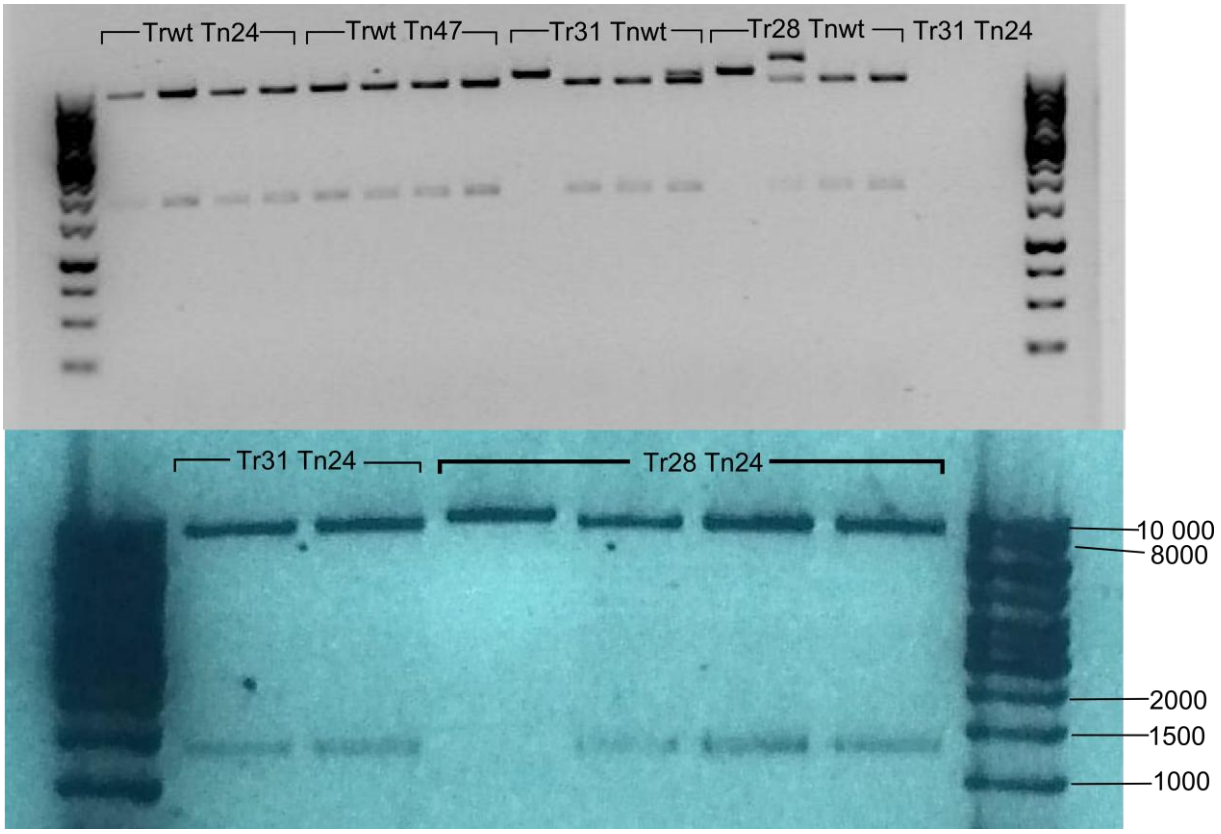

Figure 3-5: The resulting gel after separating the digested fragments. The pAO-TrTn constructs were digested with the restriction enzyme NdeI. The expected fragment pattern of a pAO-TrTn construct was a fragment of 1775 bp and a fragment of 8126 bp. Four parallels were digested and all the parallels show the expected fragment pattern except Tr31 Tnwt parallel 1, Tr28 Tnwt parallel 1 and 2, Tr31 Tn24 parallel 1 and 2 and Tr28 Tn24 parallel 1. The DNA ladder used is GeneRuler 1 Kb DNA ladder #SM0311 from Thermo Scientific (appendix E).

3.2.2 5'-UTR fusions constructed by fusing one transcription stimulating and one translation stimulating 5'-UTR can result in an additive or repressive effect on *bla* expression

To test the effect on the *bla* expression resulting from a 5'-UTR fusion consisting of a transcription stimulating and a translation stimulating region (fig. 1-4) the pLUTR plasmid was constructed (fig 2-1). The difference from the pIB11 plasmid is that the 5'-UTR in front of *bla* has been switched out with a long 5'-UTR consisting of two wt 5'-UTR sequences fused together (wt wt) with a spacer sequence containing a SpeI restriction site, allowing the replacement of the upstream and downstream 5'-UTR region. The 5'-UTR fusion variants are named in the order its components appear in the construct with the upstream region first followed by the downstream region (e.g. LV-2 Tn25 means that the 5'-UTR fusion consist of the LV-2 sequence as the upstream region and the Tn25 sequence as the downstream region). The sequences of the 5'-UTR fusions constructed in this study can be viewed in appendix F, table 8-10 and 8-11. The oligonucleotides used for the construction can be views in table 2-2.

5'-UTR Fusions combined of previously identified transcription and translation stimulating 5'-UTRs

After an initial testing of different combinations (appendix G, figure 8-3), the identified transcription stimulating 5'-UTR variant LV-2 (Berg et al. 2009) was chosen as a fixed transcription stimulating region and several different translation stimulating 5'-UTRs were fused with it, and also to the wt 5'-UTR as controls. The *bla* production of the various 5'-UTR fusions was determined by replica plating, and the results are shown in figure 3-6 along with the short corresponding 5'-UTRs for comparison.

From the results in figure 3-6 it is evident that, except for the LV-2 Tn58 variant, the 5'-UTR fusions did not result in strains with an increased ampicillin resistance compared to the strains with *bla* expression controlled by the corresponding short 5'-UTRs. Fusion with LV-2 compared to fusion with wild type as the upstream 5'-UTR region can result in increased, decreased or unchanged ampicillin resistance. The 5'-UTR fusions resulted in strains with a high ampicillin resistance under uninduced conditions, especially the LV-2 Tn58 variant leading to viable growth at 3 500 µg/mL, which was maximum ampicillin concentration for the uninduced plates. The maximum ampicillin concentration observed, when induced with 2.0 mM *m*-toluic acid, was 13 000 µg/mL. Three strains (LV-2, LV-2 Tn58 and LII-11)

showed viable growth at this concentration and are further explored later in this section (fig. 3.9).

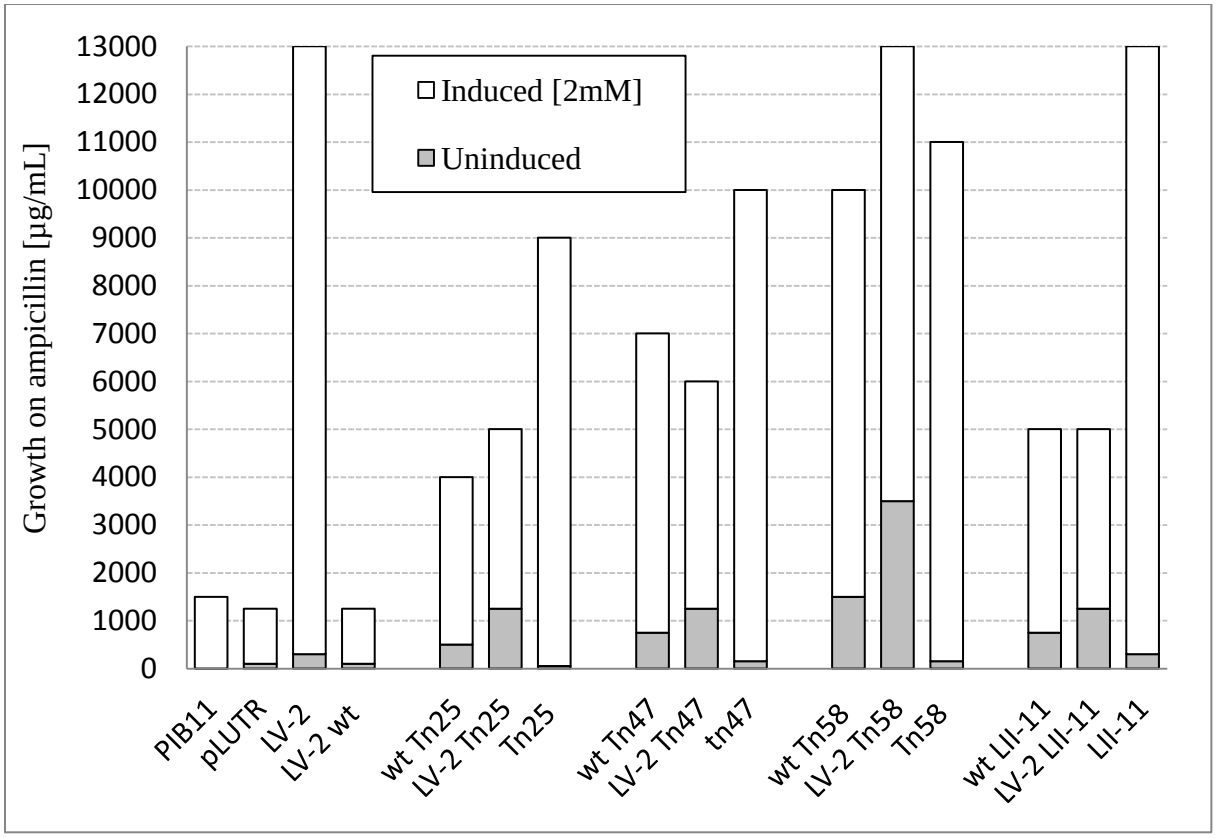

Figure 3-6: Replica plating of strains having 5'-UTR fusions in front of *bla* in the *pLUTR* construct, along with the strains having short 5'-UTRs in front of *bla* in the pIB11 construct. The two first columns are the pIB11 and pLUTR constructs without any alterations of the 5'-UTR as seen in figure 2-1. The expression was induced with 2mM *m*-toluic acid (uninduced levels are shown in grey). Maximum ampicillin concentration was 13 000 µg/mL for the induced plates and 3 500 µg/mL for the uninduced. The Wt, Tn25, pLUTR, and LV-2 wt strain had uninduced growth at 10, 50, 100 and 100 µg/mL, respectively. For complete ampicillin concentration range of the LA plates used see appendix C, table 8-4.

5'-UTR Fusions with a down variant as the downstream fusion partner

Prior to this study 5'-UTR variants causing reduced *bla* expression along with the desired trait of very low uninduced expression had been identified (Lale et al. 2011). Two of these 5'-UTR down variants (DI-3 and DI-8) were used as the downstream region in 5'-UTR fusions with LV-2 and wt in order to test if such a fusion could increase the *bla* expression while maintaining their low uninduced expression qualities. Since the down 5'-UTR variants result

in a reduced *bla* transcript amount compared to the wt 5'-UTR variant in pIB11 (Lale et al. 2011) the fusions with wt 5'-UTR might also have a stimulating effect on transcription compared to the short 5'UTR down variants. The *bla* expression was determined by replica plating (fig. 3-7).

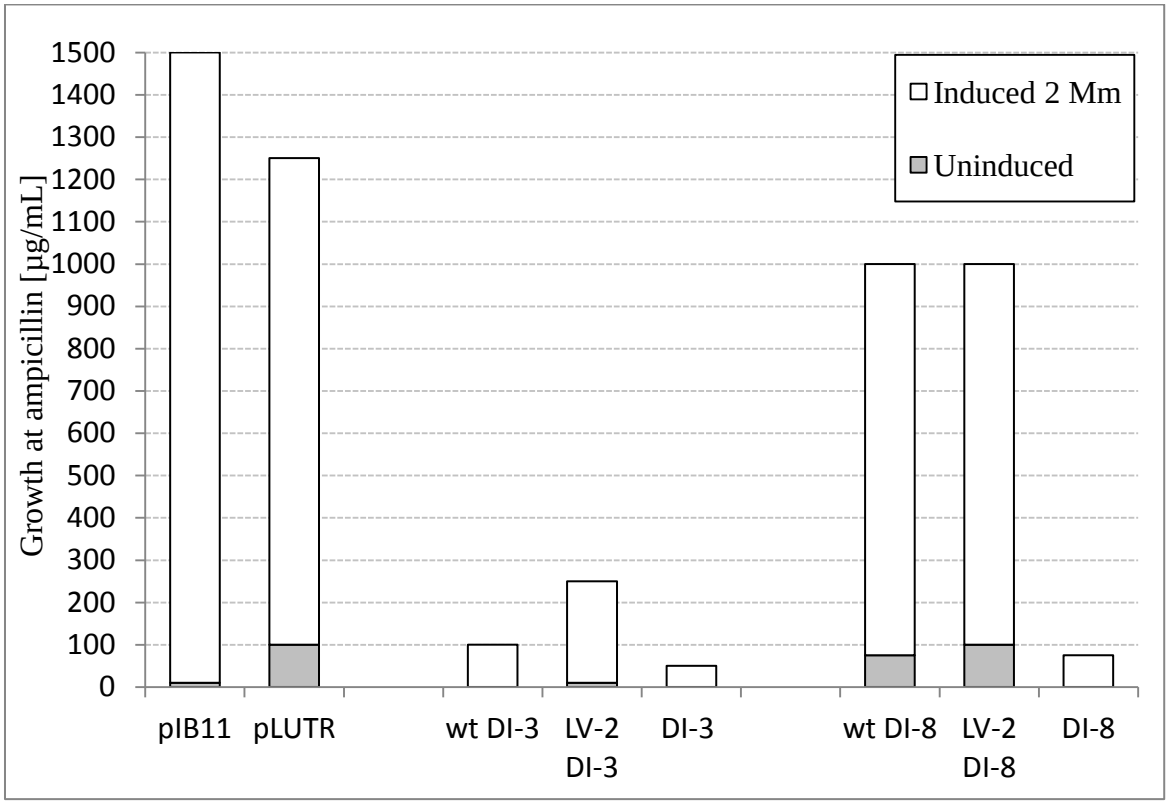

Figure 3-7: Replica plating of strains harbouring the pLUTR plasmid with the 5'-UTR down variants DI-3 and DI-8 fused with either LV-2 and wt, as well as the short DI-3 and DI-8 5'-UTRs in the pIB11 plasmid. The two first columns are the pIB11 and pLUTR constructs without any alterations of the 5'-UTR as seen in figure 2-1. The wt DI-3, DI-3 and DI-8 5'-UTR variants led to no observed growth at uninduced conditions (7,5 µg/mL was the lowest ampicillin concentration), pIB11 and LV-2 DI-3 showed growth at 10 µg/mL ampicillin.. For complete ampicillin concentration range of the plates see appendix C, table 8-4.

The DI-8 5'-UTR fusions resulted in a large increase in ampicillin resistance compared to the short DI-8 variant at both induced and uninduced conditions. The DI-3 fusions resulted in a significantly lower ampicillin resistance compared to the wt 5'-UTRs, as was expected for the down 5'-UTR variants. The LV-2 DI-3 fusion led to a 5-fold greater ampicillin resistance compared to DI-3, but the uninduced resistance had increased to the wt levels of the pIB11 strain. Interestingly the wt DI-3 variant resulted in a doubling in ampicillin resistance

compared to DI-3, while maintaining low uninduced resistance with undetected uninduced growth (lowest concentration was 7.5 µg/mL).

The sequences of the long 5’-UTR fusions were verified by sequencing as described in section 2.2.2, and the results can be viewed in appendix H, figure 8-4.

Fusions with an *in silico* made translation stimulating 5’-UTR element as the downstream fusion partner

To test how well an *in silico* made translation stimulating 5’-UTR fusion partner performs compared to those obtained through screening, the RBS calculator was used to generate 5’-UTR regions with optimal translation initiation rates (TIRs). For details on the RBS calculator and the parameters used, look to section 2.2.4. It is stated that if four 5’-UTRs are made, the probability of achieving protein expression within twofold of the target is 92% (Salis et al. 2009). Using the option for generating 5’-UTRs with maximum TIR, eight fusion partners were designed (max 1-8), four to be fused downstream of the LV-2 sequence and four to be fused downstream of the wt 5’-UTR sequence. The sequences are listed in table 3.1 along with the input and wt sequence.

Table 3-1: The 5’-UTR fusion partners, designed by the RBS calculator, (max1-8) along with the wt and input sequence. The putative SD sequence is written in bold and the restriction sites SpeI (5’-end) and NdeI (3’-end) are underscored.

Name	Sequence
input	atacactagtannnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnncatatg
wt	...........tatactagatatgtacaataataat**ggag**tcatgaa......
max 1	...........gatctaatacacccaatctttttaga**ggagg**tttta......
max 2	...........tacacgttcagcaagagcaaccgccga**ggagg**ttaa......
max 3	...........ggcgcaaccaccagttcagcaaatca**ggagg**ttctc......
max 4	...........atcccagtttacgccgaccataccctaa**ggagg**tcc......
max 5	...........gcaacaggtcctatacaactaacctaa**ggcagg**tat......
max 6	...........gcctttcagcctcagcccagaacctttaa**ggagg**ta......
max 7	...........acacgtcccccaatagttatttcttaa**ggagg**tccc......
max 8	...........ccgccactaatagtccgcgcccttaa**ggagg**aatcc......

The 5'-UTR fusions were made by cloning the max 1-4 variants downstream of LV-2 and the max 5-8 variants downstream of the wt sequence, resulting in the constructs pLUTR LV-2 max 1-4 and pLUTR wt max 5-8. The *bla* expression resulting from these 5'-UTR fusions were tested using replica plating (fig. 3-8).

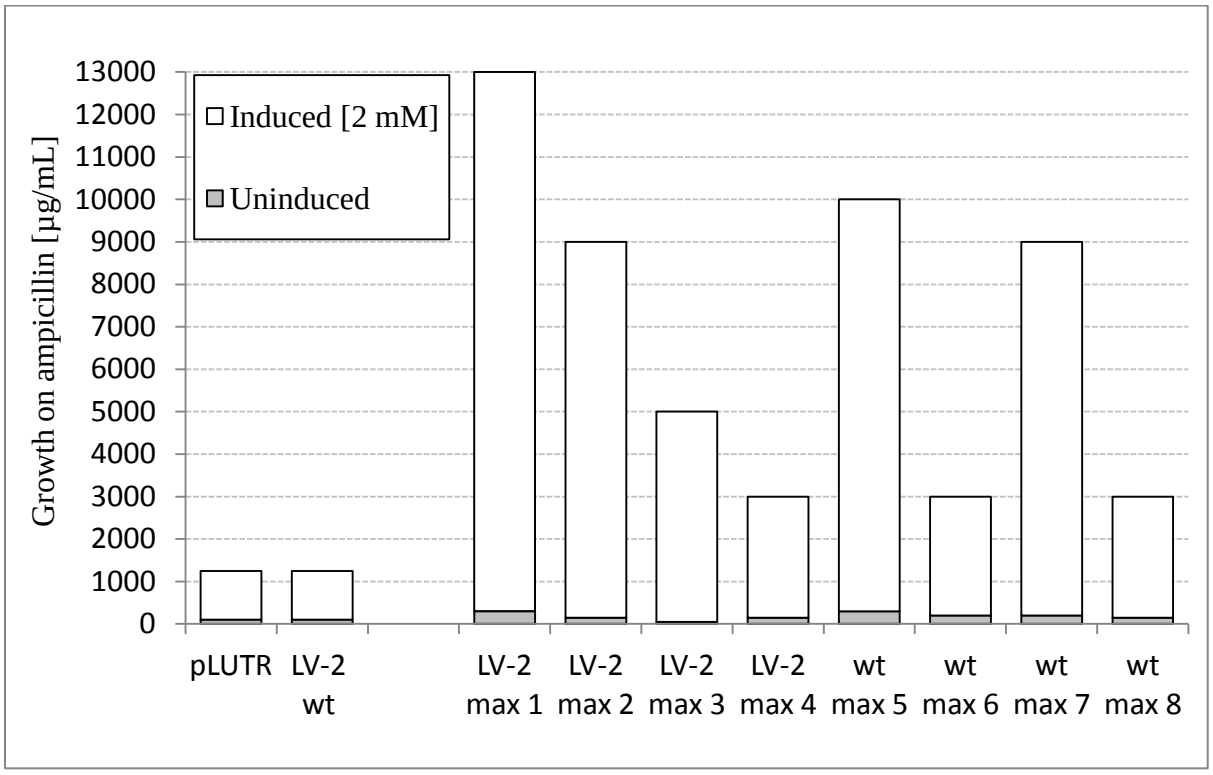

Figure 3-8: Replica plating of the strains containing 5'-UTR fusions consisting of the max sequences designed using the RBS calculator (table 3-1) fused with either the LV-2 or the wt 5'-UTR variant. The first column is the resistance level of the strain harbouring the pLUTR construct containing the wt 5'-UTR fusion sequence as seen in figure 2-1. Maximum concentration of ampicillin were 13 000 µg/mL. The LV-2 max 3 strain showed uninduced growth at 50 µg/mL. For the full range of ampicillin concentrations used see appendix C, table 8-4.

Both 5'-UTR fusions containing wt max combinations and LV-2 max combinations resulted in strains with high ampicillin tolerance (fig. 3-8). The wt max 5 strains grew on a 10 000 µg/mL ampicillin concentration, the same level as the highest wt 5'-UTR fusion with a previously identified translation stimulating fusion partner (wt Tn58, fig. 3-6), and the LV-2 max 1 strains grew at 13 000 µg/mL also achieved by the LV-2 Tn58 strain (Fig. 3-6). 13 000 µg/mL was the highest concentration used in this experiment and the strains that reached this level are further explored later in this section (fig. 3-9). The strains with an *in silico* made fusion partner gave lower ampicillin resistance at uninduced conditions compared to the long

5’-UTR strains with a previously identified translation stimulating fusion partner (fig 3-6 and 3-7).

The sequence of the long 5’-UTR max variants were verified by sequencing as described in section 2.2.2, and can be viewed in appendix H (fig. 8-5). It should be noted that the results for the wt max 5 strain showed mutations in the early *bla* sequence.

The LV-2 Tn58 5’-UTR fusion variant results in higher *bla* expression than its components on their own

The strains harbouring the pIB11 LV-2, pIB11 LII-11, pLUTR LV-2 Tn58 and pLUTR LV-2 max1 plasmids all showed viable growth on plates containing the highest concentration used of 13 000 µg/mL ampicillin, when induced with 2 mM *m*-toluic acid (fig 3-6 and 3-8). It would be impractical to make plates with a higher ampicillin concentration since too much ampicillin stock solution compared to agar will hamper the plates from solidifying. Instead one additional replica plating with lower inducer concentration (0.1 mM) was performed to differentiate the high resistance strains and identify the strain with the highest ampicillin resistance and henceforth *bla* expression (fig 3-9).

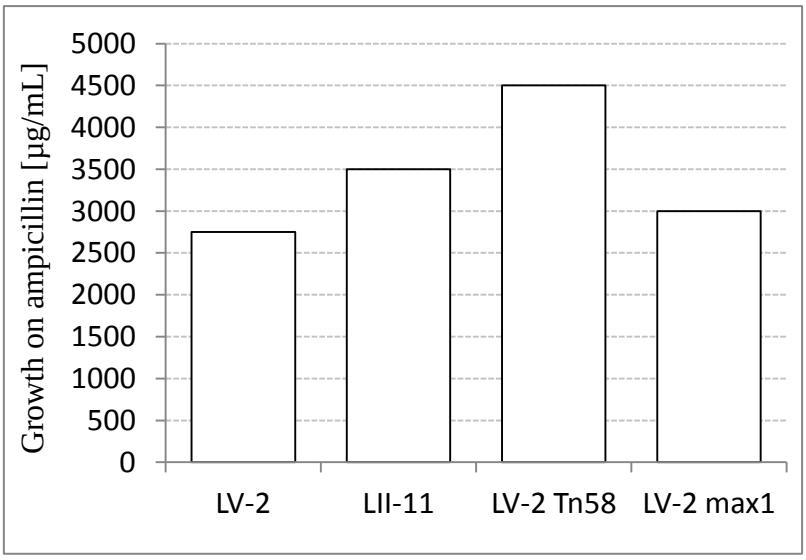

Figure 3-9: The replica plating used to differentiate the high resistance strains. Expression was induced with 0.1 mM *m*-toluic acid. For the full range of ampicillin concentrations of the plates see appendix C, table 8-5.

The highest ampicillin tolerance was observed in the strain harbouring the LV-2 Tn58 5'-UTR fusion in front of *bla* with about a 60% and 30% increase compared to LV-2 and LII-11, respectively (fig. 3-9). This result shows that the LV-2 Tn58 strain reached a higher tolerance level than what was observed for its 5'-UTR components on their own (LV-2 and Tn58).

3.3 Bioinformatics analyses of the 5'-UTR sequences discovered probable factors responsible for reduced *bla* expression

Except for the LV-2 Tn58 variant, the 5'-UTR fusions resulted in a reduced *bla* expression compared to their short 5'-UTR counterparts. To explore this phenomenon the different 5'-UTR sequences were analyzed using bioinformatics tools. The analyses explored the translation initiation rate, the structural energy of the 5'-UTR, the location of the SD sequence in relation to the secondary structures, and the strength of the SD containing secondary structures.

3.3.1 Prediction of the translation initiation rate (TIR) in respect to *bla*

Translation initiation is a crucial and often limiting step of protein synthesis. The RBS calculator predicts the rate of translation initiation with a thermodynamic model evaluating the Gibbs free energy (ΔG) of the mRNA sequence and its interaction with the 30S ribosomal complex. The TIR is given on a proportional scale from 0.001 to 100 000 + (Salis 2011).

The predicted TIRs for the short 5'-UTRs and the 5'-UTR fusions in respect to *bla* made by using the reverse engineering option in the RBS calculator (section 2.2.4) are shown in figure 3-10. The RBS calculator only takes in to account the 35 bp upstream of the start codon. Hence, it will not distinguish whether the wt or the LV-2 sequence is used as an upstream fusion partner and the difference between the fusion and the short 5'-UTRs stem from the spacer sequence percent in the 5'-UTR fusions and the removal of the PciI restriction site in the downstream 5'-UTR region of the fusions (fig 2-1).

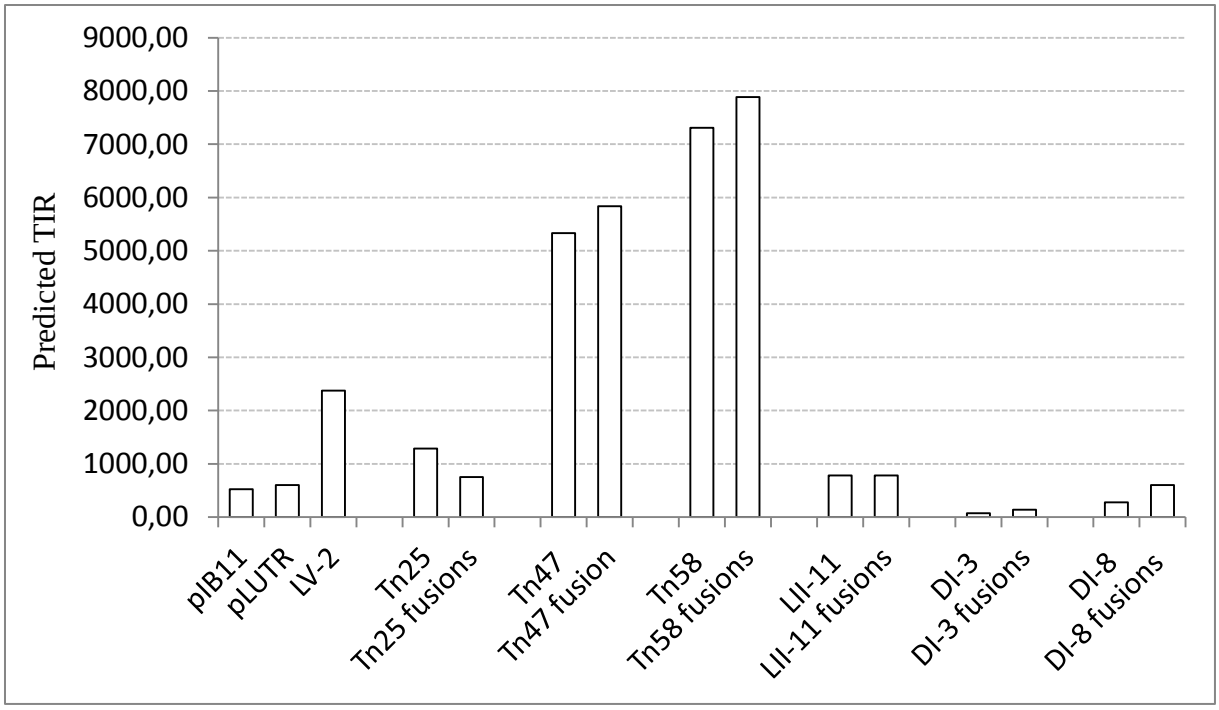

Figure 3-10: The predicted TIR for the short 5'-UTR variants together with the corresponding 5'-UTR fusions predicted by the RBS calculator. The two first columns are the TIR of the wt 5'-UTR sequences in the pIB11 and pLUTR construct, respectively, as seen in figure 2-1. All but the DI-3, DI-8 and DI-3 fusions were issued with the NEQ warning indicating that the mRNA sequence may not fold quickly to its equilibrium state (Salis 2011).

The predicted TIRs of the short 5'-UTRs (fig 3-10) show some correlation with the observed ampicillin tolerance (fig 3-6 and 3-7). The wt and down variants had a low predicted TIR and low observed ampicillin resistance. The Tn47 and Tn58 variants had a high predicted TIR and resulted in strains with high ampicillin resistance. However, the LII-11 and LV-2 sequence gave relative low predicted TIR compared to the resulting high ampicillin tolerance of the strains harbouring them. The predicted TIR for the long 5'-UTR fusions follows the same overall pattern as the TIR predictions of the short 5'-UTRs (fig. 3-10). The Tn25 fusions had a reduced predicted TIR compared to the Tn25 sequence, and the DI-8 fusions had a relatively large increase compared to the short DI-8 variant. For the remaining 5'-UTRs the predicted TIR was slightly increased or unchanged for the 5'-UTR fusions compared to the short 5'-UTRs. It should be noted that all but the DI-3 and DI-8 5'-UTR variants and the DI-3 fusions were issued with the NEQ warning indicating that the mRNA sequence may not fold quickly to its equilibrium state, which is one of the assumptions made by the RBS calculator (Salis 2011).

The predicted TIR, in respect to *bla*, of the max 1-8 5'-UTR fusion partners designed using the forward engineering option provided by the RBS calculators (tab. 3-1) are shown in figure 3.11

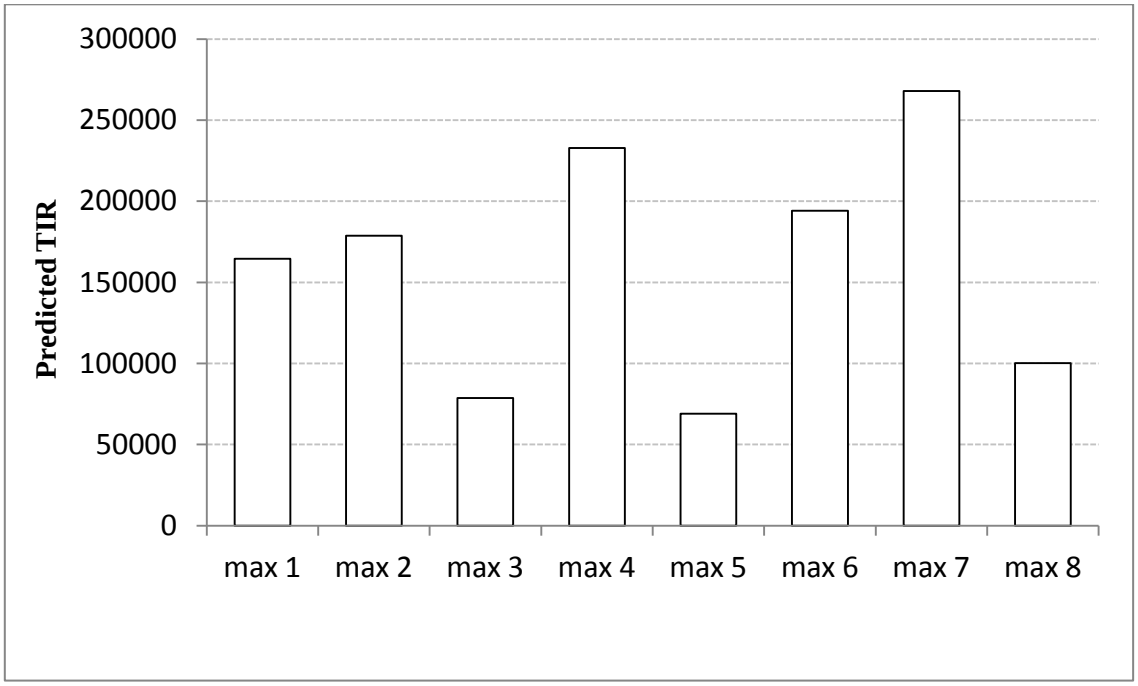

Figure 3-11: The predicted TIR for the eight 5'-UTR sequences designed using the RBS calculator (table 3-1).

The predicted TIR for the max 1-8 5'-UTR fusions (fig 3-11) showed a TIR ranging from ~ 60 000 to 260 000 all significantly higher than the TIRs predicted for the non *in silico* made sequences (fig 3-10) with a maximum TIR of ~ 7 000 – 8 000. There is no overall correlation between the predicted TIR of the max 5'-UTR sequences and the ampicillin tolerance resulting from them.

3.3.2 Evaluation of the 5'-UTR secondary structures showed an influence on *bla* expression from the structural energy and the structures at and close to the SD sequence

The secondary structures of the 5'-UTR regions were predicted using the Quickfold software as described in section 2.2.4 in order to test if some factors of the secondary structures in the 5'-UTR region reflect the observed ampicillin tolerance. The numbers of suggested structures as well as the average ΔG (kcal/mol) are summarized in appendix I (table 8-12, 8-13).

The relationship between ampicillin tolerance and the energy of the 5'-UTR secondary structures are shown in figure 3-12. The dataset is divided into the short 5'-UTRs, the wt 5'-UTR fusions and the LV-2 5'-UTR fusions.

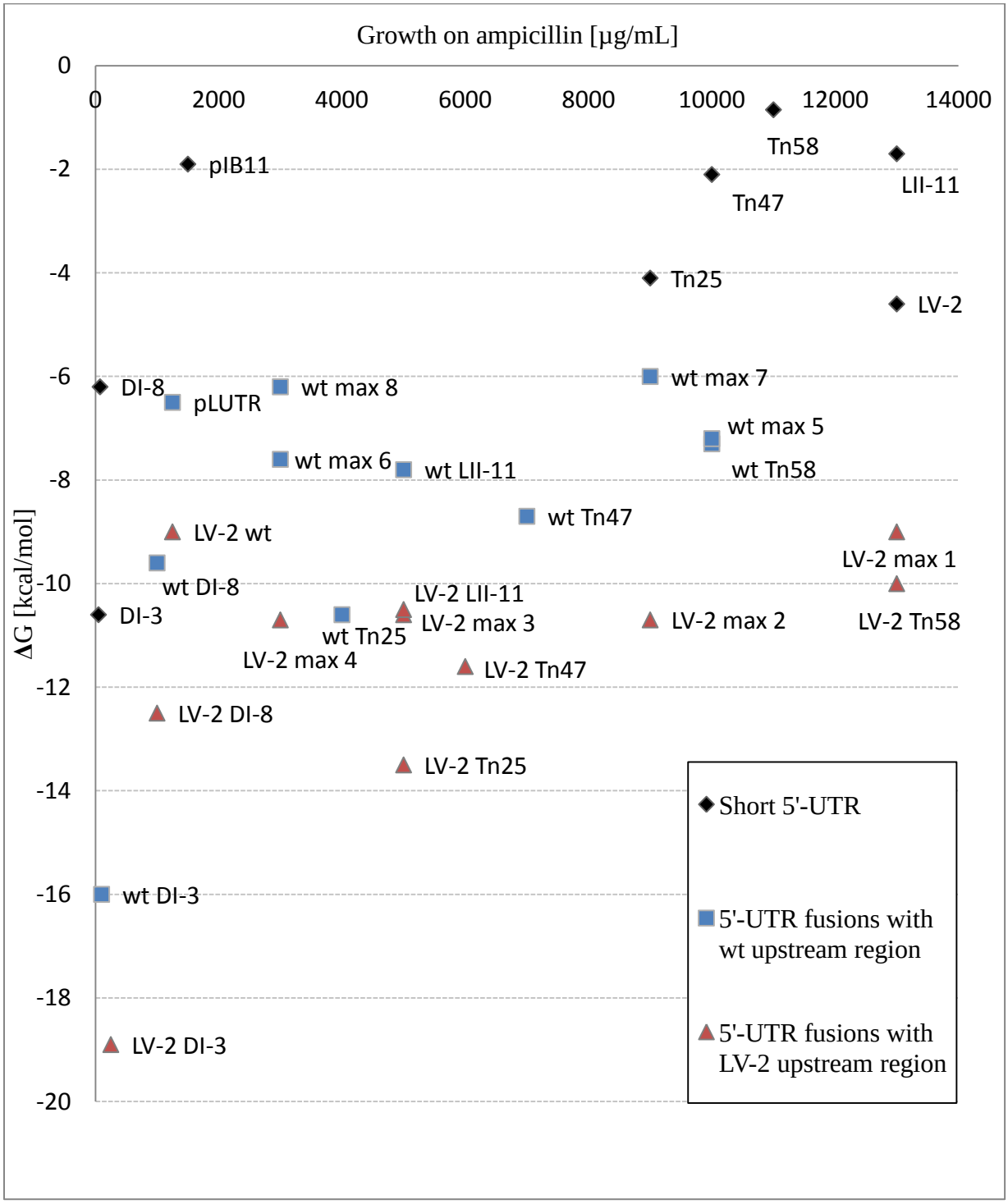

Figure 3-12: A scatter plot of the average energy of the secondary structures of the 5'-UTR regions obtained from the Quickfold software, versus the observed ampicillin resistance of the strains hosting the different 5'-UTR variants. The results are divided in to three groups; the short 5'-UTRs, the 5'-UTR wt fusions and the 5'-UTR LV-2 fusions.

From figure 3-12 it is evident that the short 5'UTRs have relatively low energy and also some of the strains with the highest ampicillin resistance. The long 5'-UTR LV2 fusions have higher structural energy than the long 5'-UTR wt fusions, probably due to the high energy of the LV-2 region (-4.6 kcal/mol) compared to the wt (at -1.9 kcal/mol). Except for the high energy of the down variants (DI-3, DI-8 and fusions containing them) which are linked with low *bla* expression, there is little correlation between energy of the secondary structure and ampicillin resistance within the groups.

To see if any differences in the location of the SD sequence in relation to the secondary structures reflect the resulting ampicillin tolerance, the predicted secondary structures were further investigated. The predicted secondary structures of the short 5'-UTRs are shown in figure 3-13. In the cases were two structures were suggested the one with the most favourable energy is shown. The structures not shown did not deviate much from the ones presented in figure 3-13.

Except from the wild type and the down variants, the secondary structures (fig 3-13) of the short 5'-UTRs have a similar shape containing two GC bonds and with the SD sequence as part of a loop at the end of a stem-loop structure. The wt, DI-3 and DI-8 have their SD as part of a more rigid structure with 3, 3, and 4 GC bonds, respectively, and the wt and DI-8 variant have the SD sequence as part of an internal loop of a stem-loop structure.

The Quickfold software suggested several different secondary structures for the 5'-UTR fusions (appendix I). Because of the low standard deviation of the energies between the suggested structures (appendix I) the amount of time the 5'-UTR sequence spends in each structure can be expected to be quite evenly distributed. An overview of the suggested structures given to each 5'-UTR fusion is given in figure 3-14. The structures are divided into four groups according to the secondary structure at and in close proximity of the SD sequence. The "free SD" group contains of structures were the putative SD sequence is not part of a secondary structure. In all the remaining three groups the SD is part of a secondary structure. However, the groups are separated by the number of additional secondary structures within 10 bp from the SD containing structure (not counting the bp in the secondary structures). The groups are exemplified in figure 3-15. The sectioning into these groups is based on the hypothesis that the more secondary structures present in close proximity of the SD sequence the more likely will they hamper the translation initiation rate.

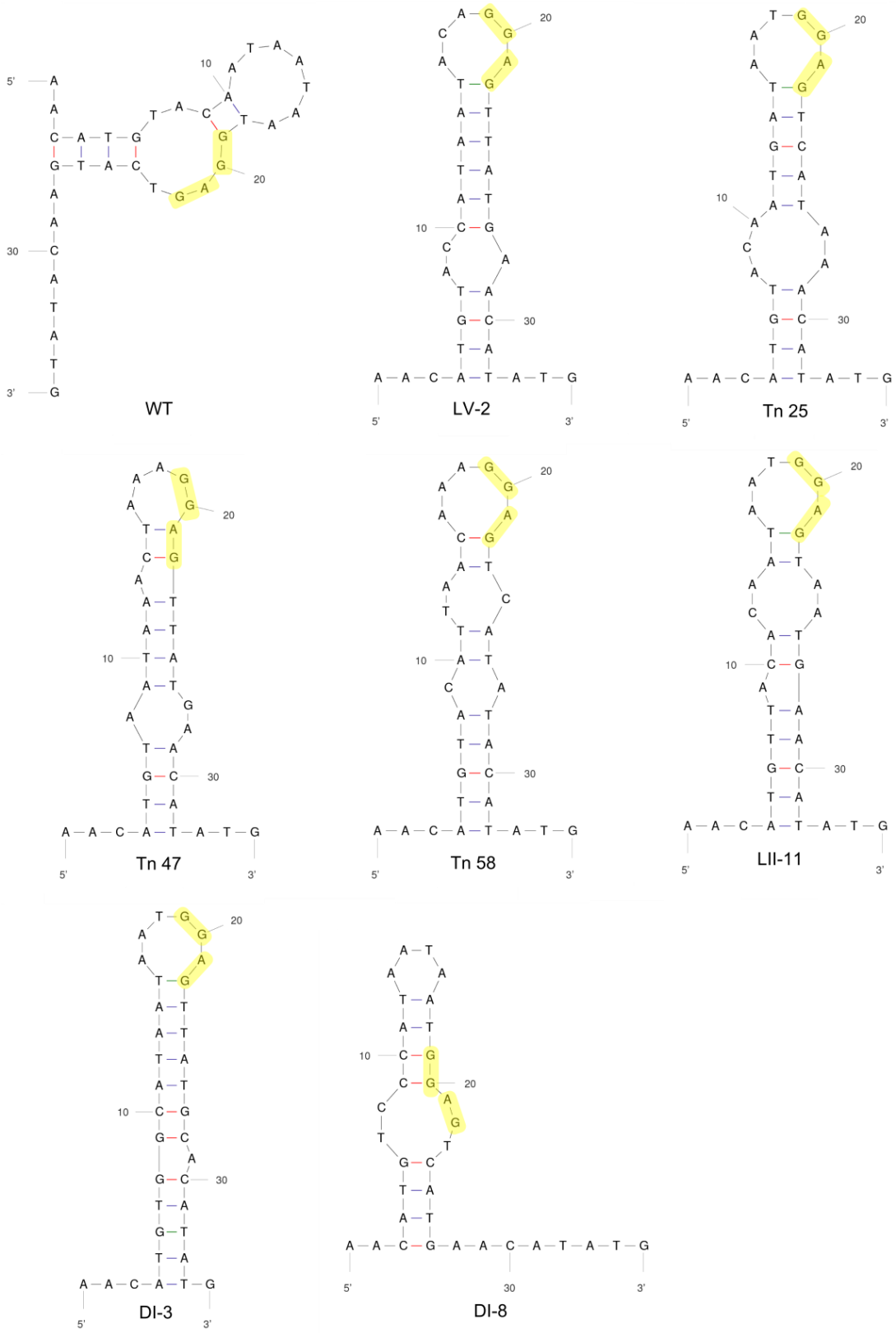

Figure 3-13: The predicted secondary structures of the short 5'-UTR sequences made by the Quickfold software. The putative SD sequences are highlighted in yellow.

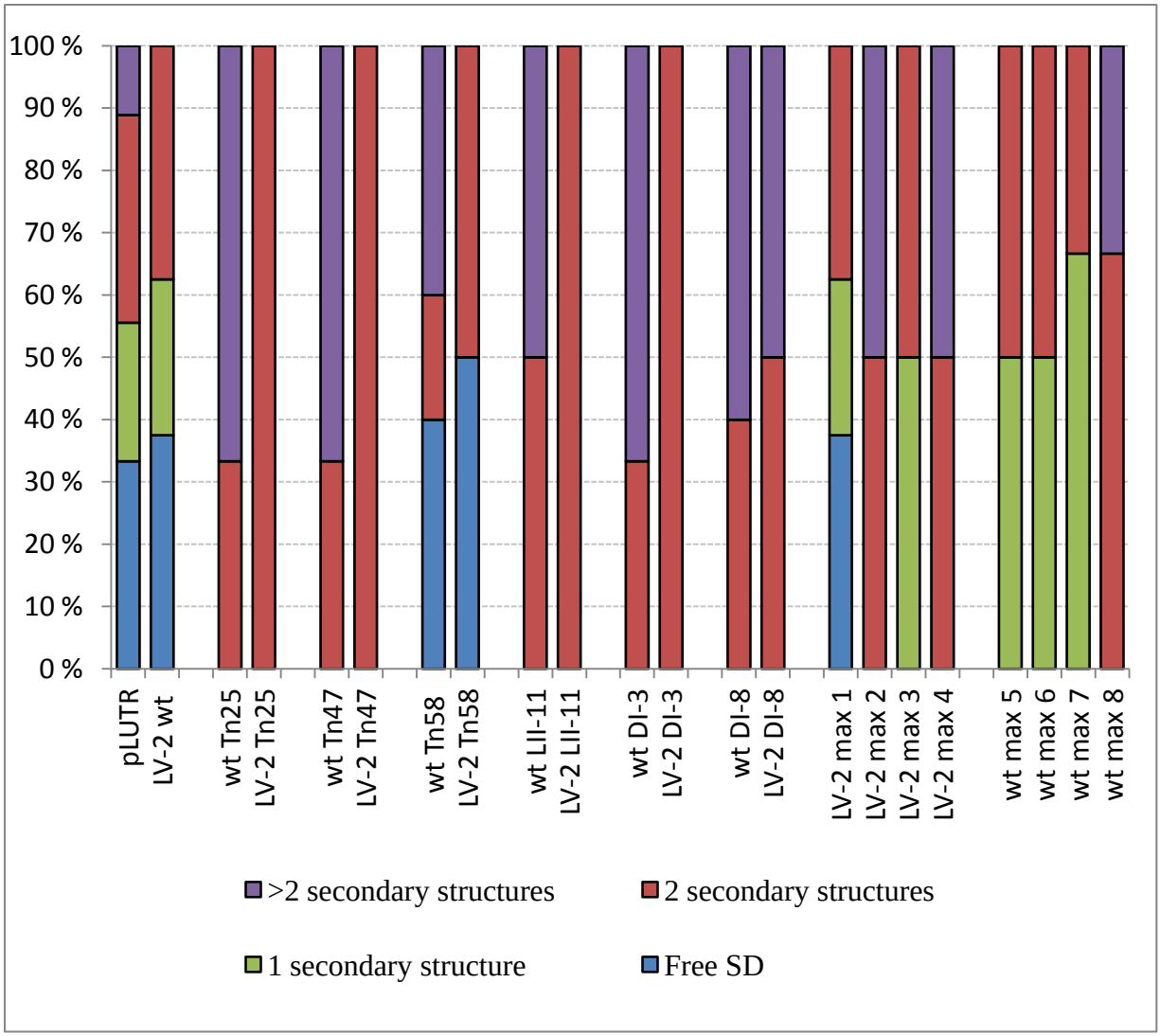

Figure 3-14: The distribution of the secondary structure groups (fig 3-15) predicted by the Quickfold software for the 5'-UTR fusions. The pLUTR column is an overview of the predicted structures of the wt 5'-UTR sequence of the pLUTR construct (wt wt) as presented in figure 2-1.

It can be observed in figure 3-14 that the 5'-UTR fusions which have a similar ampicillin resistance level or greater than their short 5'-UTR counterparts (pLUTR and LV-2 wt vs. pIB11 and the Tn58 fusions vs. Tn58) are the ones with free SD structures. The ones with decreased resistance in comparison to their short 5'-UTR counterparts (The Tn25, Tn47 and LII-11 fusions vs. Tn25, Tn47 and LII-11, respectively) all have a high proportion of 2 and > 2 secondary structures groups and no free SD structures. The down variant fusions have a high content of secondary structures and results in strains with low ampicillin resistance. Of the 5'-UTR fusions with an *in silico* made fusion partner the strain with highest ampicillin resistance is the one with free SD structures (LV-2 max 1) (fig 3-8). The ones with 2 or more

than 2 secondary structures had relatively low ampicillin resistance except for LV-2 max 2. Amongst the ones with a high proportion of predicted structures in the 1 secondary structure group; two strains showed relatively high resistance (wt max 5 and wt max 7) and two strains relatively low resistance (LV-2 max 3 and wt max 6).

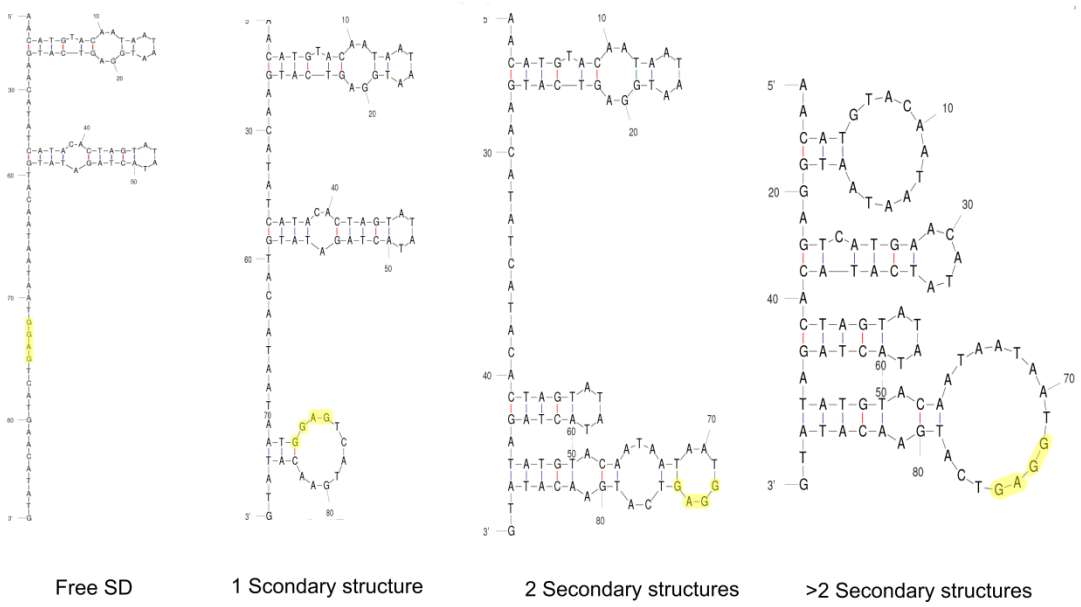

Figure 3-15: Examples of the secondary structure groups. All sequences are of the wt 5'-UTR sequence of pLUTR seen in figure 2-1. The putative SD sequence is highlighted in yellow.

3.3.3 The average content of GC bonds in the SD containing secondary structure

Bonds between guanine and cytosine are strong due to the three hydrogen bonds formed between them. Hence the number of GC bonds will reflect the strength of the secondary structure. An overview of the average GC bonds in the secondary structure containing the SD is given in figure 3-16.

The wt, Tn58 and max 1 fusion sequences have low average GC bonds in the SD containing secondary structure partly due to the high proportion of free SD structures (fig. 3-14). The other non *in silico* fusions have a two GC bond average except for the down mutants with 2.5-3 GC bonds. The max fusions show more variance, ranging from 1-5 GC bonds. The LV-2 max sequences show a correlation between low average GC bonds in the SD containing structure and high resulting ampicillin tolerance. However, for the wt max sequences some 5'-

UTR fusions led to relatively high ampicillin resistance despite high content of GC bonds in the SD containing secondary structure (wt max 5 and wt max 7).

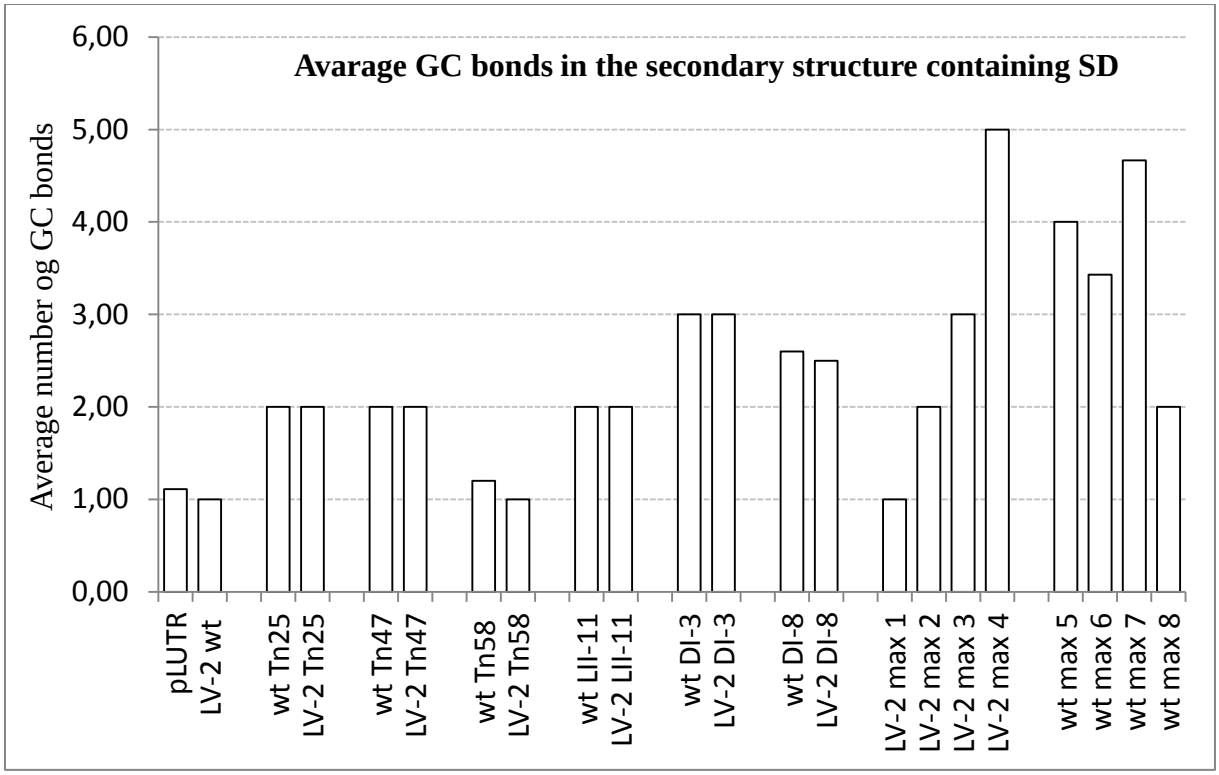

Figure 3-16: The average number of GC bonds in the secondary structure containing the SD sequence based on the predictions of the Quickfold software. Structures in the free SD group (fig 3-15) were treated as a structure with 0 GC bonds.

4 Discussion

4.1 The transcription and translation stimulating traits of the short 5'-UTRs

The results in figure 3-1 clearly indicate that the translation stimulating 5'-UTR variants (Tn) obtained through screening (see section 1.2) indeed lead to elevated ampicillin tolerance resulting from enhancement of translation. Since no increase in ampicillin tolerance was observed when the Tn 5'-UTR variant sequences were placed in the front of *celB,* influencing the transcription (fig 3-1), it seems like the Tn 5'-UTR variants does not have a stimulatory effect on transcription. However, the results from the β-lactamase assay and transcript analysis (fig. 3-3) indicate a significant increase in transcript amount compared to wt as well as increased β-lactamase activity for the Tn strains tested (Tn24 and Tn47). The two experiments differ in both construct (pAO and pIB11) and method making them unsuited for direct comparison, but the difference in apparent transcript stimulation is still striking. One possible explanation is that in the pAO double construct the transcription initiation of the *celB*-*bla* mRNA transcript are somewhat uncoupled from the translation initiation of *bla* since the transcription initiation is far upstream from the *bla* translation initiation site. Whilst in the pIB11 construct transcription and translation initiation of *bla* occurs in close proximity. Moving ribosomes prevent spontaneous backtracking by the RNAP and hence increased translation rate will lead to increased transcription rate (Proshkin et al. 2010). This effect might be less prominent in the pAO double construct since the ribosomes translating *celB* will leave the transcript before reaching *bla*. There might also be a difference in the mRNA decay rate of the resulting transcripts from the two constructs since a negative correlation between mRNA length and mRNA decay has been identified in *E. coli* (Feng et al. 2007), which will influence the transcript amount. However the mRNA decay rate was not explored in this study.

Regarding the experiment presented in figure 3-2, an increase in ampicillin tolerance is observed in both the *celB* and the *bla* position indicating that the Tr 5'-UTR variants have both transcription and translation stimulating traits. This is reflected in the β-lactamase assay and transcript analysis experiment (fig. 3-3) as well. From the results of the tested Tr 5'-UTR variants and the previously tested transcription stimulating 5'-UTRs LV-1 and LV-2 (Berg et al. 2009) it seems like a transcription stimulating 5'-UTR will also stimulate translation, even

though screening was preformed so that only *celB* expression was directly affected and an increase in *bla* expression would only be a result of an increase in transcript amount. It is possible that the traits stimulating transcription overlap to some degree with the traits stimulating translation. The transcription enhancing effect can be a result of faster promoter escape and/or faster formation of the open complex, as hypothesized by Berg et al. (Berg et al. 2009), which would explain why the effect was seen in both constructs (pIB11 and pAO).

4.2 The effect of combining a transcription stimulating and a translation stimulating 5'-UTR in the pAO-TrTn construct

The combination of transcription stimulating 5'-UTR in front of *celB* and a translation stimulating 5'-UTRs in front of *bla* in the pAO-TrTn construct (fig. 1-3 and 2-2) led to a higher ampicillin tolerance than the stimulating 5'-UTR variants did in combination with the wt 5'-UTR (fig- 3-4). Due to the design of the pAO construct (section 1.2) the tolerance increase probably results from the combination of both increasing the *celB-bla* mRNA transcript amount as well as the translation of *bla*. Analysis of the differences in transcript amount and mRNA decay rate would give more information on the factors resulting in the increase.

4.3 A putative promoter element in the spacer region may explain the observed high uninduced *bla* expression for some 5'-UTR fusions

A striking trait of these long UTR fusions was a substantial increase in ampicillin resistance during uninduced conditions compared to their short 5'-UTR counterparts (fig. 3-6, 3-7). However, this increase was not observed in the 5'-UTR fusions with an *in silico* made downstream fusion partner (fig. 3-8). This ongoing trend indicates that there is some factor differing from the *in silico* fusions responsible. That might be the spacer sequence downstream of the SpeI restriction site and/or a part of the conserved downstream 5'-UTR sequence (fig. 2-2). One hypothesis is that the presence of a promoter sequence in the 5'-UTR fusion may lead to the additional uninduced gene expression. Indeed, a potential -10 promoter element with five of six nucleotides corresponding to the consensus promoter sequence exists in the spacer downstream of the SpeI restriction site which could act as a functional promoter together with a possible -35 promoter element 15 bp upstream (17 bp being the optimal

distance) with four of six nucleotides corresponding to the consensus sequence (for location of the possible promoter elements see appendix J, figure 8-6). Based on the position of the putative -10 element a transcript from this promoter would contain a 5'-UTR region close to the length of the short 5'-UTR variants, thus having the RBS intact. Amongst the 5'-UTR fusions with high uninduced ampicillin the LV-2 fusions had a higher uninduced expression compared to the corresponding wt fusions (fig 3-6). It might be that some of the mutations in the LV-2 sequence favour the activity of the putative promoter. The -10 element is not present in the *in silico* made fusion (tab. 3-1, fig. 3-8) which might explain why they did not result in strains with as high uninduced ampicillin tolerance.

4.4 Remarks on the bioinformatics analyses performed on the 5'-UTR fusions

To shed light on the properties affecting the gene expression of the 5'-UTR fusions several bioinformatics analyses were performed on the 5'-UTR sequences (section 3.3). The RBS calculator calculates the Translation initiation rate (TIR) by evaluating the Gibbs free energy (ΔG) of the sequence and its interaction with the ribosome (fig. 3-10). Obviously more factors than the TIR will determine the gene expression, for instance the LV-2 5'-UTR variant leads to high *bla* expression but still had a relatively low TIR prediction. This is probably because part of the high expression caused by the LV-2 5'-UTR variant comes from transcription stimulation. Still some 5'-UTR sequences like LII-11 and Tn 25 gave lower predicted TIRs than one would expect based on their ampicillin tolerance. Of course all prediction tools take in to account a limited set of factors and it should be noted that most of the sequences subjected to reverse engineering got the "NEQ" warning indicating that a model assumption might not be fulfilled (Salis 2011).

The free energy (ΔG) of the most favourable secondary structures of the 5'-UTR sequences was predicted by the Quickfold software. As can be visualized in figure 3-12 the short 5'-UTRs had on average a less stable structure than the long 5'-UTR fusions and amongst the fusions the LV-2 combinations had the most stable structures. The increased stability of the LV-2 fusions compared to the wt fusions can be explained by that the LV-2 5'-UTR sequence on its own had a higher structural energy with -4.6 kcal/mol as opposed to -1.9 kcal/mol for the wt 5'-UTR. The down variants and the 5'-UTR fusions containing them had high structural stability accompanied with low gene expression, especially for the DI-3 5'-UTR

variant and DI-3 fusions. Apart from the down variants no clear relationship between the 5'-UTR stability and gene expression was observed within the groups. It is known that stable structures in the mRNA 5'-end will lower the gene expression (Kudla et al. 2009; Tuller et al. 2010b). The structural energy, however, tends to have less of an impact on the gene expression when approaching lower levels (Supek et al. 2010; Tuller et al. 2010b). This may explain why only the most stable structures showed a convincing correlation with low gene expression. It is possible that the short 5'-UTRs are more sensitive towards high stability since they contain fewer nucleotides to distribute the energy, and that is why low ampicillin tolerance is observed for DI-3 and DI-8 at energy levels of -10.6 and -6.2 kcal/mol, respectively. The 5'-UTR fusions seem to withstand a higher stability, possibly because the secondary structures will distribute over a larger area, and some structures will be further away from the SD sequence and start codon and hence might have less of an impact on the translation initiation.

The predicted secondary structures of the short 5'-UTRs were quite similar (fig. 3-13) as one would expect since the sequences themselves are very similar and the short stretch of nucleotides limits the number of possible structures. Most of the structures had the putative SD-sequence as part of an end-loop in a stem-loop structure containing two GC bonds. The exceptions are the ones resulting in the lowest ampicillin resistance. The wt and DI-8 5'-UTR sequence had the SD sequence as part of an internal loop in a stem-loop structure and the wt, DI-3 and DI-8 had stronger structures with 3, 3 and four GC bonds, respectively. It is possible that a SD sequence in an internal loop is less accessible by the ribosome leading to hampering of the translation. The same goes for a high degree of GC bonds which makes for a stronger secondary structure.

Most of the 5'-UTR fusions gave a large number of predicted secondary structures. To get a better overview, the secondary structures were divided into groups depending on the secondary structure at and close to the SD sequence (fig. 3-14). The sectioning into groups was based on the hypothesis that the more secondary structures present close to the SD sequence the more hindrance of the translation will occur. In addition the amount of GC bonds in the SD containing secondary structure was analyzed because the strength of the GC bonds contributes to a more stable structure. A SD sequence within a strong secondary structure will be less accessible which might hamper translation.

The 5'-UTR fusions with unstructured SD sequence (free SD) consist of the strains which achieved about the same ampicillin tolerance level or higher than the corresponding short 5'-UTRs (pLUTR, LV-2 wt, wt Tn58 and LV-2 Tn58), and also the two 5'-UTR fusions resulting in the highest ampicillin tolerance (LV-2 Tn58 and LV-2 max 1). In addition the 5'-UTR fusions with 2 and >2 secondary structures were associated with reduced (compared to short UTRs) and low gene expression (with the exception of LV-2 max2). This shows that valuable information not picked up by just looking at the sequence stability can be obtained by closer inspection of the predicted secondary structures.

The analyses performed in this study are more or less linked together. The RBS calculator model is based on the mRNA stability (Δ G) which again is affected by the number of GC bonds in the SD containing structure. 5'-UTRs with structures in the free SD group will automatically have a lower average number of GC bonds in the SD containing secondary structure. However the results show that they are different enough so that one analysis can pick up information overlooked by another.

4.5 Limiting factors discovered can explain low/reduced ampicillin tolerance for the 5'-UTR fusions

The wt 5'-UTR sequence of pLUTR and the LV-2 wt 5'-UTR fusion resulted in strains with close to the same ampicillin tolerance as the short wt 5'-UTR sequence of pIB11 (fig.3-6). The analysis showed a low TIR (fig. 3-10), relatively low mRNA stability (fig. 3-12), structures with a free SD sequence (fig. 3-14) and a low average of GC bindings in the SD containing secondary structures (fig. 3-16). The three last traits are believed to favour high gene expression so the results indicate that some unfavourable interaction with the ribosome picked up by the RBS calculator is the limiting factor (low TIR) for the strain with the short and long wt 5'-UTR, as well as the LV-2 wt fusion.

The wt Tn25 and LV-2 Tn25 5'-UTR fusions resulted in a lower ampicillin tolerance with growth at 4 000 and 5 000 µg/mL ampicillin, respectively, compared to 9 000 µg/mL resulting from the Tn25 5'-UTR variant. The Tn25 fusions had a lower predicted TIR than the Tn25, relatively high Δ G and highly structured sequences close to the SD sequence. Wt Tn25 had more structures in the >2 secondary structures category than LV-2 Tn25 which might explain its lower tolerance level. Alternatively it is expected that the LV-2 fusions produce

more transcript. The results indicate that the decrease in tolerance level compared to the Tn25 strain comes from a decrease in TIR and/or highly structured 5'-UTRs with relatively high stability.

The wt Tn47 and LV-2 Tn47 resulted in strains with relatively high ampicillin resistance with growth at 7 000 and 6 000 µg/mL ampicillin, respectively, but still substantial lower than the Tn47 strain with viable growth at 10 000 µg/mL. The Tn47 sequence has a relatively high predicted TIR and the Tn47 fusions had a slight increase, suggesting that the TIR is not the limiting factor. The 5'-UTR stability (Δ G) does not stand out as especially high compared to the other 5'-UTR fusions, but the Tn47 combinations have at least two secondary structures at and close to the SD sequence. The same goes for the LII-11 fusions resulting in a substantial decrease in ampicillin tolerance with growth at 5 000 µg/mL ampicillin down from more than 13 000 for the LII-11 strain despite an unchanged predicted TIR compared to LII-11 and Δ G close to the ones predicted for the 5'-UTR fusions with the highest ampicillin tolerance. So in the case of the Tn47 and LII-11 5'-UTR fusions it seems like the structures at and close to the SD sequence is the limiting factor reducing the *bla* expression.

The wt Tn58 and LV-2 Tn58 stood out with a resulting ampicillin tolerance close to and surpassing that of Tn58 with growth at 10 000 and more than 13 000 µg/mL ampicillin respectively, compared to 11 000 µg/mL for the Tn58 strain. It is noteworthy that these two sequences got favourable results in all the factors analyzed. They have the highest predicted TIR surpassing that of Tn58, the second lowest Δ G next to the wt combinations and high proportion of secondary structures with "free SD" resulting in a low GC bond average of the SD containing secondary structure. The higher resistance level resulting from LV-2 Tn58 compared to wt Tn58 might come from increased transcription due to stimulation from the LV-2 sequence, which could be further investigated with transcript analysis. In addition the LV-2 Tn58 5'-UTR variant has more favourable predicted secondary structures than the wt Tn58 5'-UTR with a higher proportion of free SD and no structures in the >2 secondary structures group. It should also be noted that the LV-2 Tn58 5'-UTR resulted in an especially high *bla* expression at uninduced conditions with growth at more than 3500 µg/mL. If this uninduced expression is due to the putative promoter in the spacer region (as discussed above) it is possible that it will contribute to an increased *bla* expression also during induced condition since repetitive promoters elements in tandem have been shown to increase gene expression (Li et al. 2012).

The DI-3 fusions resulted in an increase in ampicillin tolerance of about 2-fold for the wt DI-3 strain and 5-fold for the LV-2 DI-3 strain compared to the strain harbouring the short DI-3 5'-UTR (fig. 3-7). The wt DI-3 fusion looks to have maintained the low uninduced expression level of the DI-3 strain with undetected growth at an ampicillin concentration of 7.5 µg/mL, while the LV-2 DI-3 strain showed uninduced growth at the levels reached by the wt short 5'-UTR. The DI-3 fusions share traits with the DI-3, with the lowest predicted TIR, the highest 5'-UTR stability and equal amount of GC bonds in the SD containing secondary structure. Since the transcript amount of the DI-3 is about half that of the wt strain (Lale et al. 2011) it is possible that the 2-fold increase in expression from the wt DI-3 fusion is a result of increased transcription resulted from the wt 5'-UTR fusion partner. Increased transcription is of course also a possibility for the LV-2 DI-3 due to the transcription stimulating properties of the LV-2 sequence. In addition the LV-2 DI-3 5'-UTR does not have any predicted secondary structures in the >2 secondary structures group as opposed to wt DI-3 which might contribute to the higher expression level. The DI-3 fusions did not result in such an increase of expression at uninduced conditions as observed in the other 5'-UTRs containing the -10 promoter element. It might be that some part of the DI-3 sequence has a negative effect on the transcription from that promoter region.

The DI-8 fusions both resulted in about a 13-fold increase in ampicillin tolerance compared to the DI-8 strain. The DI-8 fusions had an increased TIR to about wt levels and average GC bonds in the SD containing sequence of about 2.5 down from 4 in the DI-8 5'-UTR. These traits may explain the increase in ampicillin resistance as well as the possibility of increased transcript amount. Like the other 5'-UTR fusions the uninduced tolerance levels were high, possibly because of the -10 promoter element present in the spacer region.

The max 5'-UTR variants (tab. 3-1), designed by the RBS calculator, are based on the maximizing of the TIR value alone. Hence it was no surprise that the predicted TIR of the max 5'-UTR sequences were much higher than that of the other sequences studied (fig. 3-10). However, the substantial difference was not reflected in the experimentally observed ampicillin resistance where they performed quite evenly with the 5'-UTR fusions composed of sequences obtained by selection screening (fig. 3-6 and 3-8). There is little correlation between the TIR and the observed ampicillin resistances of the max fusions. This, in combination with that all the predicted TIRs for the max fusions are much higher than the other 5'-UTR fusions with greater ampicillin resistance, indicate that other factors than the

TIR is the limiting factor in determining the gene expression in the case of the max fusion strains.

The LV-2 max 1 5'-UTR resulted in the highest ampicillin resistance level of all the max fusions only surpassed by the LII-11 and LV-2 Tn*58* strain. Like the LV-2 Tn*58* 5'-UTR the LV-2 max 1 5'-UTR performed favourably in all the categories analyzed with the lowest 5'-UTR stability of all the LV-2 fusions together with LV-2 wt, and predicted secondary structures with a free SD sequence. The LV-2 max 2 strain had viable growth at the relatively high ampicillin concentration of 9 000 µg/mL. It has a high number of predicted secondary structures at and close to the SD sequence. However, a low average number of GC bonds in the SD containing structure of two may explain the relatively high ampicillin tolerance. The LV-2 max 3 5'-UTR resulted in a medium ampicillin tolerance with growth at 5 000 µg/mL ampicillin. Except for the DI-3 fusions, the LV-2 max 3 strain had the lowest background expression with no observed growth at 75 µg/mL ampicillin. It has predicted secondary structures with only one structure at the SD sequence. However, with an average of 3 GC bonds it is quite stable which might be the limiting factor. The LV-2 max 4 strain showed relatively low *bla* expression with growth at 3 000 µg/mL ampicillin, which might be explained by the high number of secondary structures at and close to the SD sequence in combination with 5 GC bonds in the SD containing structure.

The wt max 5 strain showed a relative high ampicillin resistance with viable growth at 10 000 µg/mL ampicillin, the same level as the wt Tn*58* strain, making these two strains the wt fusions with the highest *bla* expression. The strain has a high *bla* expression despite a stable SD containing structure with an average of 4 GC bonds. However, half of the predicted structures belonged in the 1 secondary structure group supporting good accessibility of the SD sequence. A theory which seeks to explain why high expression can be observed despite stable structures covering the SD sequence is named "the stand by model" (Unoson et al. 2007) which is based on that the ribosomes are able to bind a single stranded part of the mRNA close to the SD sequence, thus be in immediate presence when the SD sequence is accessible. In the case of wt max 5 the sequence in close proximity of the SD is indeed single stranded. The same goes for the wt max 7 strain which share much of the traits with wt max 5 and also results in a relatively high gene expression. In addition it has a quite low Δ G with -6.2 kcal/mol. Wt max 6 as well as wt max 5 and 7 have predicted secondary structures in the 1 secondary structure group combined with a high number of GC bonds in the SD containing structure. However, it results in low ampicillin tolerance indicating some differences not

picked up by the analyses performed in this study. The wt max 8 strain too had relatively low ampicillin resistance probably because of many structures close to the SD sequence. It should be noted that the wt max 5 strain showed mutations in the initial *bla* sequence (appendix G). It has been shown that the nucleotides in the initial coding region can have a significant impact on gene expression (Ahn et al. 2008; Tuller et al. 2010b)

4.6 5'-UTR fusions enhancing transcription and translation may lead to increased/high *bla* expression when limiting factors are absent

It is noteworthy that the two 5'-UTR fusions (LV-2 Tn58 and LV-2 max 1) that resulted in the highest ampicillin tolerance performed favourably in all the categories analyzed with no obvious limiting factor emerging. It should also be noted that these two 5'-UTR fusions are LV-2 fusions. No ongoing trend linked high expression with a LV-2 fusion partner, but it is a possibility that the absence of other limiting factors have allowed the transcription stimulating properties of LV-2 to contributed to the high *bla* expression for these two strains. No data on the transcript amount was obtained for the 5'-UTR fusions in this study. Thus, it is not known whether the LV-2 sequence has kept its transcription stimulating properties as part of a 5'-UTR fusion. However, considering that the transcriptional open complex is only 12-15 bp long, initially just covering a couple of bases after the TSS and that the RNAP usually have formed the elongation complex by the time it has reached the nucleotide +20 of the transcript mRNA (DeHaseth et al. 1998; Davis et al. 2007) it seems unlikely that the sequence downstream of the LV-2 region (stating at +35) of the 5'-UTR fusions will have a significant influence on transcription initiation. Since the LV-2 5'-UTR variant has a confirmed transcript enhancing effect in the pIB11 construct (Berg et al. 2009), there is a good chance it has the same effect in the pLUTR construct as well. For most of the other LV-2 5'-UTR fusions the analyses picked up one or more possible limiting factors which might overshadow the transcription stimulating effect from the LV-2 region.

When further investigating the strains with the highest ampicillin resistance the LV-2 Tn58 strain was found to have the highest tolerance, close to 30% greater than that resulted from the LI-11 strain (fig. 3-9) previously shown to have an increase in β-lactamase activity of a factor of 20 compared to the pIB11 wt strain (Berg et al. 2009). Hence, it reached a higher ampicillin resistance level than both of its components (LV-2 and Tn58). The LV-2 max 1 strain barely surpassed the ampicillin tolerance of the LV-2 strain previously shown to have

about a 17 fold increase of *bla* expression compared to the pIB11 wt strain (Berg et al. 2009). This indicates that the *in silico* designed 5'-UTR sequences are capable of reaching the same expression levels as those identified by selective screening.

5 Conclusion

The aim of this study was to look into the possibility of designing a 5'-UTR fusion consisting of a transcriptional stimulating and a translational stimulating region capable of achieving higher gene expression than what was achieved by its components alone. The results indicate that the creation of a 5'-UTR fusions tend to introduce limiting factors yielding a reduced gene expression. However some 5'-UTR fusions successfully resulted in high gene expression and the LV-2 Tn58 strains surpassed the ampicillin resistance level of both its components (the LV-2 and Tn58 strain) showing a possible additive effect of stimulating both transcription and translation in the form of a 5'-UTR fusion. This indicates that testing a relatively small number of different sequences gives a good chance of success. The method also seemed viable to increase expression of down variants while maintaining low uninduced expression.

The results indicate that *in silico* designed translation stimulating 5'-UTR sequences on average are able to perform at the same levels as the 5'-UTR sequences obtained trough selection screening. The LV-2 max 1 variant showed that these 5'-UTR fusions have the potential of reaching *bla* expression levels on par with the previously identified high expressive 5'-UTRs.

It seems likely that the high gene expression seen from some of the LV-2 fusions is due to increased transcription stimulation caused by the LV-2 region, though without data on the transcript amount a conclusion cannot be made.

6 Future prospects

Transcript analysis such as described in section 2.2.3 should be performed on the 5'-UTR fusions to confirm if the LV-2 sequence stimulate transcription as part of a 5'-UTR fusion in such a way as it does on its own in the pIB11 construct (Berg et al. 2009). Analysis of any differences in mRNA decay could be useful to investigate if it might explain some of the variance in expression. To get a more accurate measurement on the variances in gene expression (as opposed to replica plating) protein assays should be performed as described in section 2.2.3.

To investigate the possible transcription from the -10 promoter element found in the spacer region one could seek to analyse the transcript length since the resulting transcript would be shorter than the transcript from the *Pm* promoter. Alternate spacer regions can be tested and if the high uninduced expression disappears, as seen in the max fusions, it would support that an active promoter element is responsible. In addition testing the effect of alternate spacer regions would be interesting regardless of the investigation of the promoter element.

It would be interesting to further test the robustness of the analyses performed on the 5'-UTR fusions by designing new 5'-UTR fusions with high TIR, unstructured SD sequence and low structural energy, to see if these also would result in strains with high expression of the target gene. Other reporter genes could be tested to investigate to what degree the observed effect of the 5'-UTR fusions is gene specific.

Making a 5'-UTR library with a fixed LV-2 upstream region and a randomized downstream region could be performed to see if it would identify 5'-UTR regions outperforming the ones covered in this study.

7 References

Ahn, J. H., J. W. Keum, et al. (2008). "High-throughput, combinatorial engineering of initial codons for tunable expression of recombinant proteins." Journal of Proteome Research **7**(5): 2107-2113.

Arraiano, C. M., J. M. Andrade, et al. (2010). "The critical role of RNA processing and degradation in the control of gene expression." FEMS Microbiology Reviews **34**(5): 883-923.

Aune, T. E. V. and F. L. Aachmann (2010). "Methodologies to increase the transformation efficiencies and the range of bacteria that can be transformed." Applied Microbiology and Biotechnology **85**(5): 1301-1313.

Babitzke, P., C. S. Baker, et al. (2009). "Regulation of translation initiation by RNA binding proteins." Annual Review of Microbiology **63**: 27-44.

Berg, L., V. Kucharova, et al. (2012). "Exploring the 5'-UTR DNA region as a target for optimizing recombinant gene expression from the strong and inducible Pm promoter in *Escherichia coli*." J Biotechnol **158**(4): 224-230.

Berg, L., R. Lale, et al. (2009). "The expression of recombinant genes in *Escherichia coli* can be strongly stimulated at the transcript production level by mutating the DNA-region corresponding to the 5'-untranslated part of mRNA." Microb Biotechnol **2**(3): 379-389.

Bethesda (1986). "BRL pUC Host: *E. coli* DH5a Competent cells " Focus **vol. 8** (2): 13.

Bio-Rad "Bio-Rad Protein Assay." 1-27.

Biolabs, N. E. "Double Digest Finder ". from http://www.neb.com/nebecomm/DoubleDigestCalculator.asp#.T8DQzZIpXmx.

Blatny, J. M., T. Brautaset, et al. (1997). "Construction and use of a versatile set of broad-host-range cloning and expression vectors based on the RK2 replicon." Appl Environ Microbiol **63**(2): 370-379.

Blattner, F. R., G. Plunkett Iii, et al. (1997). "The complete genome sequence of *Escherichia coli* K-12." Science **277**(5331): 1453-1462.

Bradford, M. M. (1976). "A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein dye binding." Analytical Biochemistry **72**(1-2): 248-254.

Breaker, R. (2011). "Prospects for Riboswitch Discovery and Analysis." Molecular Cell **43**(6): 867-879.

Browning, D. F. and S. J. W. Busby (2004). "The regulation of bacterial transcription initiation." Nature Reviews Microbiology **2**(1): 57-65.

Burger, A., C. Whiteley, et al. (2011). "Current perspectives of the *Escherichia coli* RNA degradosome." Biotechnology Letters **33**(12): 2337-2350.

Cèbe, R. and M. Geiser (2006). "Rapid and easy thermodynamic optimization of the 5′-end of mRNA dramatically increases the level of wild type protein expression in *Escherichia coli*." Protein Expression and Purification **45**(2): 374-380.

Chen, H., M. Bjerknes, et al. (1994). "Determination of the optimal aligned spacing between the Shine-Dalgarno sequence and the translation initiation codon of *Escherichia coli* mRNAs." Nucleic Acids Research **22**(23): 4953-4957.

Davis, C. A., C. A. Bingman, et al. (2007). "Real-time footprinting of DNA in the first kinetically significant intermediate in open complex formation by *Escherichia coli* RNA polymerase." Proceedings of the National Academy of Sciences of the United States of America **104**(19): 7833-7838.

DeHaseth, P. L., M. L. Zupancic, et al. (1998). "RNA polymerase-promoter interactions: The comings and goings of RNA polymerase." J Bacteriol **180**(12): 3019-3025.

dos Reis, M., L. Wernisch, et al. (2003). "Unexpected correlations between gene expression and codon usage bias from microarray data for the whole *Escherichia coli* K-12 genome." Nucleic Acids Research **31**(23): 6976-6985.

Dougan, D. A., A. Mogk, et al. (2002). "Protein folding and degradation in bacteria: To degrade or not to degrade? That is the question." Cellular and Molecular Life Sciences **59**(10): 1607-1616.

Erie, D. A. (2002). "The many conformational states of RNA polymerase elongation complexes and their roles in the regulation of transcription." Biochimica et Biophysica Acta - Gene Structure and Expression **1577**(2): 224-239.

Feng, L. and D. K. Niu (2007). "Relationship between mRNA stability and length: An old question with a new twist." Biochemical Genetics **45**(1-2): 131-137.

Goldman, S. R., R. H. Ebright, et al. (2009). "Direct detection of abortive RNA transcripts in vivo." Science **324**(5929): 927-928.

Huang, C. J., H. Lin, et al. (2012). "Industrial production of recombinant therapeutics in *Escherichia coli* and its recent advancements." Journal of Industrial Microbiology and Biotechnology **39**(3): 383-399.

Jana, S. and J. K. Deb (2005). "Strategies for efficient production of heterologous proteins in *Escherichia coli*." Applied Microbiology and Biotechnology **67**(3): 289-298.

Jung, S. T., T. H. Kang, et al. (2010). "Efficient expression and purification of human aglycosylated Fcγ receptors in *Escherichia coli*." Biotechnology and Bioengineering **107**(1): 21-30.

Kaczanowska, M. and M. Rydén-Aulin (2007). "Ribosome biogenesis and the translation process in *Escherichia coli*." Microbiology and Molecular Biology Reviews **71**(3): 477-494.

Kortmann, J. and F. Narberhaus (2012). "Bacterial RNA thermometers: Molecular zippers and switches." Nature Reviews Microbiology **10**(4): 255-265.

Kudla, G., A. W. Murray, et al. (2009). "Coding-Sequence Determinants of Expression in *Escherichia coli*." Science **324**(5924): 255-258.

Lale, R., L. Berg, et al. (2011). "Continuous control of the flow in biochemical pathways through 5' untranslated region sequence modifications in mRNA expressed from the broad-host-range promoter Pm." Appl Environ Microbiol **77**(8): 2648-2655.

Laursen, B. S., H. P. Sørensen, et al. (2005). "Initiation of protein synthesis in bacteria." Microbiology and Molecular Biology Reviews **69**(1): 101-123.

Lehman, I. R. (1974). "DNA ligase: structure, mechanism, and function." Science **186**(4166): 790-797.

Li, M., J. Wang, et al. (2012). "A strategy of gene overexpression based on tandem repetitive promoters in *Escherichia coli*." Microbial Cell Factories **11**(19): 1-10.

Ma, J., A. Campbell, et al. (2002). "Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures." J Bacteriol **184**(20): 5733-5745.

Nudler, E. and M. E. Gottesman (2002). "Transcription termination and anti-termination in *E. coli*." Genes to Cells **7**(8): 755-768.

Perri, S., D. R. Helinski, et al. (1991). "Interactions of plasmid-encoded replication initiation proteins with the origin of DNA replication in the broad host range plasmid RK2." Journal of Biological Chemistry **266**(19): 12536-12543.

Pingoud, A. and A. Jeltsch (2001). "Structure and function of type II restriction endonucleases." Nucleic Acids Research **29**(18): 3705-3727.

Proshkin, S., A. Rachid Rahmouni, et al. (2010). "Cooperation between translating ribosomes and RNA polymerase in transcription elongation." Science **328**(5977): 504-508.

Ramos, J. L., S. Marqués, et al. (1997). Transcriptional control of the *Pseudomonas* TOL plasmid catabolic operons is achieved through an interplay of host factors and plasmid-encoded regulators. **51:** 341-373.

Richardson, C. C. (1965). "Phosphorylation of nucleic acid by an enzyme from T4 bacteriophage-infected Escherichia coli." Proceedings of the National Academy of Sciences of the United States of America **54**(1): 158-165.

Roberts, R. J. (2005). "How restriction enzymes became the workhorses of molecular biology." Proceedings of the National Academy of Sciences of the United States of America **102**(17): 5905-5908.

Saecker, R. M., M. T. Record Jr, et al. (2011). "Mechanism of bacterial transcription initiation: RNA polymerase - Promoter binding, isomerization to initiation-competent open complexes, and initiation of RNA synthesis." Journal of Molecular Biology **412**(5): 754-771.

Salis, H. M. (2011). "The ribosome binding site calculator." Methods in Enzymology **498**: 19-42.

Salis, H. M., E. A. Mirsky, et al. (2009). "Automated design of synthetic ribosome binding sites to control protein expression." Nature Biotechnology **27**(10): 946-950.

Schweder, T., K. H. O. Lee, et al. (1996). "Regulation of *Escherichia coli* starvation sigma factor (σs) by ClpXP protease." J Bacteriol **178**(2): 470-476.

Sharp, P. M. and W. H. Li (1987). "The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications." Nucleic Acids Res **15**(3): 1281-1295.

Simonetti, A., S. Marzi, et al. (2009). "A structural view of translation initiation in bacteria." Cellular and Molecular Life Sciences **66**(3): 423-436.

Sletta, H., A. Nedal, et al. (2004). "Broad-host-range plasmid pJB658 can be used for industrial-level production of a secreted host-toxic single-chain antibody fragment in *Escherichia coli*." Appl Environ Microbiol **70**(12): 7033-7039.

Storz, G., J. Vogel, et al. (2011). "Regulation by Small RNAs in Bacteria: Expanding Frontiers." Molecular Cell **43**(6): 880-891.

Supek, F. and T. Šmuc (2010). "On relevance of codon usage to expression of synthetic and natural genes in *Escherichia coli*." Genetics **185**(3): 1129-1134.

Taylor, F. J. R. and D. Coates (1989). "The code within the codons." BioSystems **22**(3): 177-187.

Tuller, T., A. Carmi, et al. (2010a). "An evolutionarily conserved mechanism for controlling the efficiency of protein translation." Cell **141**(2): 344-354.

Tuller, T., Y. Y. Waldman, et al. (2010b). "Translation efficiency is determined by both codon bias and folding energy." Proceedings of the National Academy of Sciences of the United States of America **107**(8): 3645-3650.

Unoson, C. and E. G. H. Wagner (2007). "Dealing with stable structures at ribosome binding sites: Bacterial translation and ribosome standby." RNA Biology **4**(3): 113-117.

Villalobos, A., J. E. Ness, et al. (2006). "Gene Designer: a synthetic biology tool for constructing artificial DNA segments." BMC Bioinformatics **7**(285): 1-8.

Vimberg, V., A. Tats, et al. (2007). "Translation initiation region sequence preferences in *Escherichia coli*." BMC Molecular Biology **8**(100): 1-13.

Waley, S. G. (1974). "A spectrophotometric assay of beta-lactamase action on penicillins." Biochem J **139**(3): 789-790.

Walter, S. and J. Buchner (2002). "Molecular chaperones - Cellular machines for protein folding." Angewandte Chemie - International Edition **41**(7): 1099-1113.

Welch, M., S. Govindarajan, et al. (2009). "Design parameters to control synthetic gene expression in *Eschorichia coli*." PLoS One **4**(9): 1-10.

Winther-Larsen, H. C., J. M. Blatny, et al. (2000). "Pm promoter expression mutants and their use in broad-host-range RK2 plasmid vectors." Metabolic Engineering **2**(2): 92-103.

Zipper, H., H. Brunner, et al. (2004). "Investigations on DNA intercalation and surface binding by SYBR Green I, its structure determination and methodological implications." Nucleic Acids Res **32**(12): e103.

Zuker, M. (2003). "Mfold web server for nucleic acid folding and hybridization prediction." Nucleic Acids Research **31**(13): 3406-3415.

8 Appendices

A. Annealing program

Table 8-1: The settings used on the PCR machine when annealing oligonucleotides

Step	Temperature (ºC)	Time (min)
1	95	10
2	80	2
3	75	2
4	70	3
5	65	5
6	55	10
7	50	10
8	45	3
9	40	2
10	35	2
11	30	1
12	25	1
13	20	1
14	15	1
15	4	Hold

B. The *bla* sequence used as input in the RBS calculator

50 fist nucleotides of the *bla* DNA sequence:

ATGAGTATTCAACATTTTCGTGTCGCCCTTATTCCCTTTTTTGCGGCATTTTG

C. Ampicillin concentrations of the plates used in replica plating

Table 8-2: The ampicillin concentration of the plates used in the replica plating presented in figure 3-1 and 3-2.

	Uninduced plates	Induced plates (0.5 mM *m*-toluic acid)
Plate	Ampicillin concentration [µg/mL]	Ampicillin concentration [µg/mL]
1	5	50
2	7.5	100
3	10	200
4	25	300
5	50	400
6	100	500
7	150	750
8	200	1000
9	250	2000
10	300	3000
11	350	4000
12	400	5000
13		6000

Table 8-3: The ampicillin concentration of the plates used in the replicatplating presented in figure 3-4.

	Uninduced plates	Induced plates (0.5 mM *m*-toluic acid)
Plate	Ampicillin concentration [µg/mL]	Ampicillin concentration [µg/mL]
1	2.5	50
2	5	100
3	7.5	250
4	10	500
5	25	750
6	50	1000
7	100	1500
8	250	2000
9	500	3000
10	1000	4000
11		5000
12		6000
13		7000
14		8000

Table 8-4: The ampicillin concentration of the plates used in the replicatplating presented in figure 3-6, 3-7 and 3-8.

	Plates without inducer	Plates with inducer (2.0 mM *m*-toluic acid)
Plate	Ampicillin concentration [µg/mL]	Ampicillin concentration [µg/mL]
1	7.5	25
2	10	50
3	25	75
4	50	100
5	75	250
6	100	500
7	150	750
8	200	1000
9	250	1250
10	300	1500
11	400	2000
12	500	3000
13	750	4000
14	1000	5000
15	1250	6000
16	1500	7000
17	2000	8000
18	2500	9000
19	3000	10000
20	3500	11000
21		12000
22		13000

Table 8-5: The ampicillin concentration of the plates used in the replikaplating presented in figure 3-9

	Induced plates (0.1 mM *m*-toluic acid)
Plates	Ampicillin concentration [µg/mL]
1	1000
2	1250
3	1500
4	2000
5	2500
6	2750
7	3000
8	3250
9	3500
10	4000
11	4500
12	5000
13	6000

D. β-lactamase assay and qRT-PCR data

Table 8-6: The slopes of the linear area of the graphs obtained when measuring the absorbance as for the β-lactamase assay as described in section 2.2.3.

Sample	Parallel 1	Parallel 2	Parallel 3	Average	St DEV	% St DEV
pB11 wt	-8.621	-7.084	-6.844	-7.516	0.964	12.828
pB11 LV-1	-93.7	-106.6	-111.2	-103.833	9.072	8.737
pB11 n24	-71.18	-72.24	-76.98	-73.467	3.088	4.204
pB11 n47	-81.92	-89.06	-82.1	-84.360	4.071	4.826
pB11 r28	-61.21	-56.08	-65.61	-60.967	4.770	7.823
pB11 r31	-30.19	-30.32	-31.94	-30.817	0.975	3.164

Table 8-7: The absorbance measurements of known protein (BSA) concentrations performed to make the protein concentration standard curve (fig. 8.1).

BSA protein Concentration (mg/mL)	Parallel 1	Parallel 2	Parallel 3	Average	St DEV
0	0.3036	0.3063	0.3224	0.3108	0.0102
0.05	0.42	0.4294	0.4271	0.4255	0.0049
0.1	0.4537	0.5403	0.5879	0.5273	0.0680
0.2	0.7176	0.7764	0.7785	0.7575	0.0346
0.3	0.862	0.9885	0.9671	0.9392	0.0677
0.4	1.1577	1.0801	1.1861	1.1413	0.0549

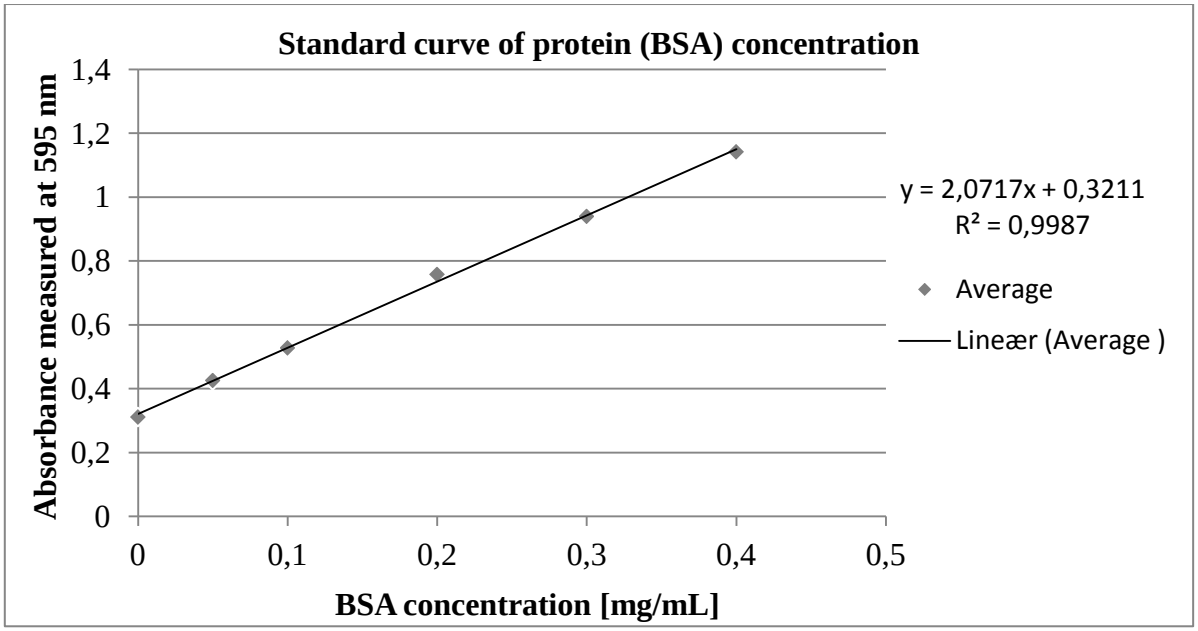

Figure 8-1: The standard curve of protein amount used to calculate the protein concentration of the samples seen in figure 3-3.

Table 8-8: Calculated protein amount of the samples presented in figure 3-3 by using the standard curve in figure 8-1.

Sample	Parallel 1	Parallel 2	Parallel 3	Average	ST DEV	Concentration (mg/mL)
pB11 wt	0.7146	0.7298	0.7422	0.7289	0.0138	0.1968
pB11 LV-1	0.6982	0.7097	0.7356	0.7145	0.0191	0.1898
pB11 n24	0.732	0.6867	0.7180	0.7122	0.0231	0.1888
pB11 n47	0.6505	0.6776	0.7489	0.6923	0.0508	0.1792
pB11 r28	0.6379	0.6625	0.7236	0.6747	0.0441	0.1707
pB11 r31	0.6903	0.7186	0.7468	0.7186	0.0282	0.1919

Table 8-9: Output data from the qRT-PCR used to determine the transcript amount presented in figure 3-3. The empty parallel boxes are omitted outliners.

Sample	Parallel 1	Parallel 2	Parallel 3	Average Ct	Ct st dev			
16S rRNA								
pIB11	-	11.987	12.275	12.131	0.144			
LV-1	-	13.000	13.010	13.005	0.005			
Tn24	13.638	13.429	-	13.534	0.105			
Tn47	14.588	14.523	14.478	14.530	0.032			
Tr28	14.377	14.094	14.208	14.226	0.082			
Tr31	12.499	12.214	12.405	12.372	0.084			
bla						RQ	RQ min	RQ max
pIB11	22.000	22.180	-	22.090	0.090	1.000	0.603	1.660
LV-1	19.604	19.739	19.684	19.676	0.039	9.769	8.955	10.657
Tn24	21.676	-	21.417	21.547	0.129	3.852	2.344	6.330
Tn47	21.435	21.326	21.434	21.398	0.036	8.517	7.760	9.347
Tr28	21.370	21.190	21.280	21.280	0.052	7.491	6.213	9.032
Tr31	20.887	20.826	20.757	20.824	0.038	2.843	2.383	3.393

E. GeneRuler 1 Kb DNA ladder #SM0311

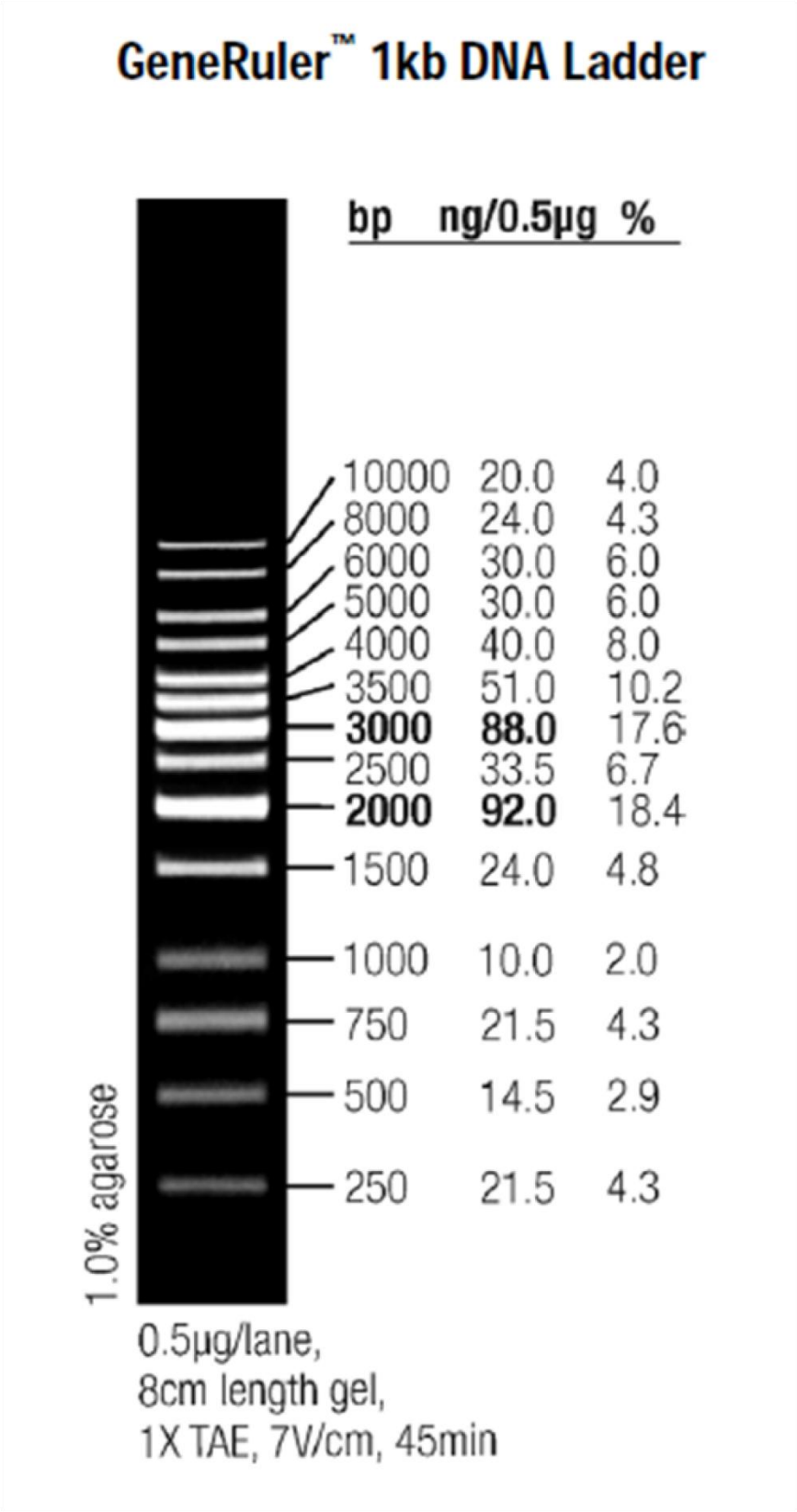

Figure 8-2: The GeneRuler 1 Kb DNA ladder #SM0311 used in the gel electrophoreses in figure 3-5

F. 5’-UTR fusion sequences

Table 8-10: The sequence of the 5’-UTR fusions presented in figure 3-6 and 3-7. The restriction sites; PciI (5’-end), SpeI (spacer) and NdeI (3’-end) are underscored and the SD sequence is marked in bold. Mutations in relation to the wt sequence are marked in red.

pLUTR	5’-AACATGTACAATAATAATGGAGTCATGAACATATCATACACTAGTATATACTAGATATGTACAATAATAAT**GGAG**TCATGAACATAG-3’
LV-2 wt	5’-AACATGTACCATAATACAGGAGTTATGAACATATCATACACTAGTATATACTAGATATGTACAATAATAAT**GGAG**TCATGAACATATG-3’
wt Tn25	5’-AACATGTACAATAATAATGGAGTCATGAACATATCATACACTAGTATATACTAGATATGTACAATGATAAT**GGAG**TCATAAACATATG-3’
LV-2 Tn25	5’-AACATGTACCATAATACAGGAGTTATGAACATATCATACACTAGTATATACTAGATATGTACAATGATAAT**GGAG**TCATAAACATATG-3’
wt Tn47	5’-AACATGTACAATAATAATGGAGTCATGAACATATCATACACTAGTATATACTAGATATGTAATAAACTAAA**GGAG**TTATGAACATATG-3’
LV-2 Tn47	5’-AACATGTACCATAATACAGGAGTTATGAACATATCATACACTAGTATATACTAGATATGTAATAAACTAAA**GGAG**TTATGAACATATG-3’
wt Tn58	5’-AACATGTACAATAATAATGGAGTCATGAACATATCATACACTAGTATATACTAGATATGTACATTAACAAA**GGAG**TCATATACATATG-3’
LV-2 Tn58	5’-AACATGTACCATAATACAGGAGTTATGAACATATCATACACTAGTATATACTAGATATGTACATTAACAAA**GGAG**TCATATACATATG-3’
wt DII-11	5’-AACATGTACAATAATAATGGAGTCATGAACATATCATACACTAGTATATACTAGATATGTTACACAATAAT**GGAG**TAATGAACATATG-3’
LV-2 DI-11	5’-AACATGTACCATAATACAGGAGTTATGAACATATCATACACTAGTATATACTAGATATGTTACACAATAAT**GGAG**TAATGAACATATG-3’
wt DI-3	5’-AACATGTACAATAATAATGGAGTCATGAACATATCATACACTAGTATATACTAGATATGTGGCATAATAAT**GGAG**TTATGCACATATG-3’
LV-2 DI-3	5’-AACATGTACCATAATACAGGAGTTATGAACATATCATACACTAGTATATACTAGATATGTGGCATAATAAT**GGAG**TTATGCACATATG-3’
wt DI-8	5’-AACATGTACAATAATAATGGAGTCATGAACATATCATACACTAGTATATACTAGATATGTCCCATAATAAT**GGAG**TCATGAACATATG-3’
LV-2 DI-8	5’-AACATGTACCATAATACAGGAGTTATGAACATATCATACACTAGTATATACTAGATATGTCCCATAATAAT**GGAG**TCATGAACATATG-3’

Table 8-11: sequence of the 5'-UTR fusions presented in figure3-8. The restriction sites; PciI (5'-end), SpeI (spacer) and NdeI (3'-end) are underscored and the SD sequence is marked in bold.

LV-2 max 1	5'-AACATGTACCATAATACAGGAGTTATGAACATATCATACACTAGTAGATCTAATACACCCAATCTTTTTAG**AGGAGG**TTTTACATATG-3'
LV-2 max 2	5'-AACATGTACCATAATACAGGAGTTATGAACATATCATACACTAGTATACACGTTCAGCAAGAGCAACCGCCG**AGGAGG**TTAACATATG-3'
LV-2 max 3	5'-AACATGTACCATAATACAGGAGTTATGAACATATCATACACTAGTAGGCGCAACCACCAGTTCAGCAAATC**AGGAGG**TTCTCCATATG-3'
LV-2 max 4	5'-AACATGTACCATAATACAGGAGTTATGAACATATCATACACTAGTAATCCCAGTTTACGCCGACCATACCCTA**AGGAGG**TCCCATATG-3'
wt max 5	5'-AACATGTACAATAATAATGGAGTCATGAACATATCATACACTAGTAGCAACAGGTCCTATACAACTAACCTA**AGGCAGG**TATCATATG-3'
wt max 6	5'-AACATGTACAATAATAATGGAGTCATGAACATATCATACACTAGTAGCCTTTCAGCCTCAGCCCAGAACCTTTA**AGGAGG**TACATATG-3'
wt max 7	5'-AACATGTACAATAATAATGGAGTCATGAACATATCATACACTAGTAACACGTCCCCCAATAGTTATTTCTTA**AGGAGG**TCCCCATATG-3'
wt max 8	5'-AACATGTACAATAATAATGGAGTCATGAACATATCATACACTAGTACCGCCACTAATAGTCCGCGCCCTTA**AGGAGG**AATCCCATATG-3'

G. Initial replica plating

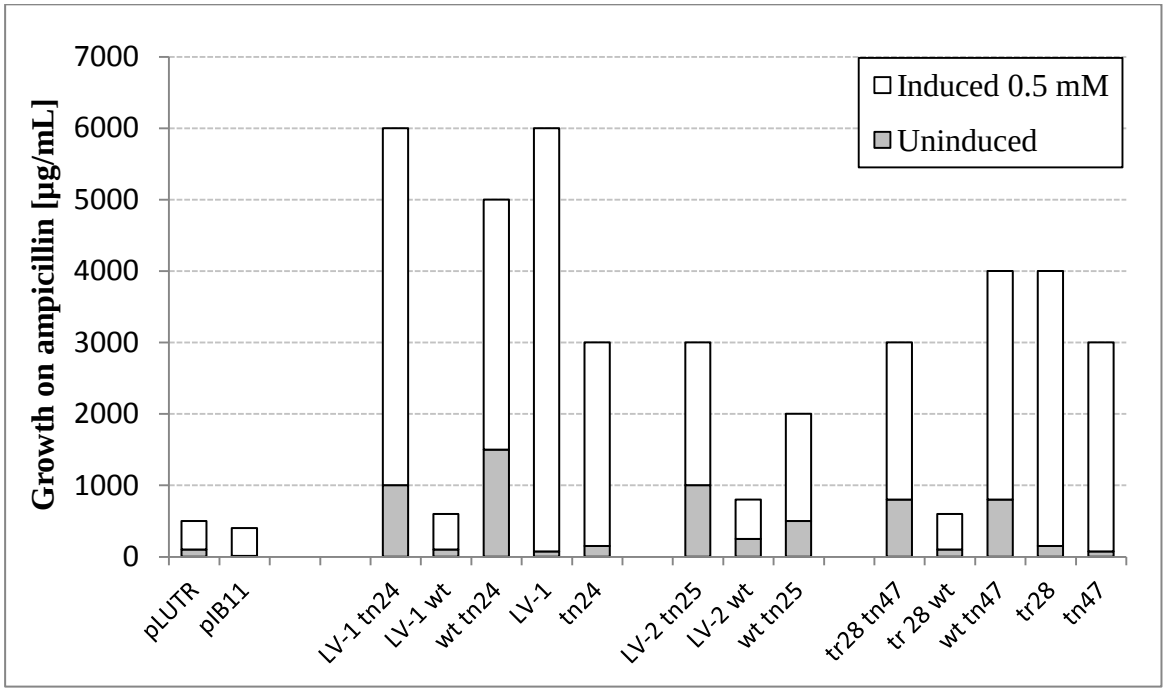

Figure 8-3: Initial replica plating performed prior to the replikaplating presented in figure 3-6 and 3-7.

H. Sequencing results

Figure 8-4: Aligned sequencing results of the 5'-UTR fusions with previously identified translation 5'-UTR variants as a downstream fusion partner

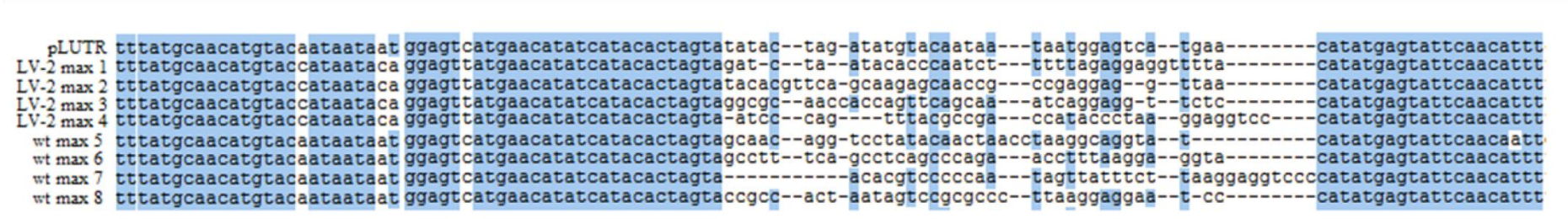

Figure 8-5: Aligned sequencing results of the 5'-UTR fusions with an in silico made downstream partner

I. Secondary structures predicted by the Quickfold software

Table 8-12: The number of predicted secondary structures for the 5'-UTR fusion sequences given by the Quickfold software, as well as the average free energy (ΔG) including standard deviation.

Strain	Number of suggested structures	Average ΔG [kcal/mol] + standard deviation
pLUTR	9	-6.5 ± 0.4
LV-2 wt	8	-9 ± 0.4
wt Tn25	3	-10.6 ± 0.4
LV-2 Tn25	1	-13.5 ± 0.0
wt Tn47	3	-8.7 ± 0.4
LV-2 Tn47	1	-11.6 ±0.0
wt Tn58	5	-7.3 ± 0.4
LV-2 Tn58	2	-10 ± 0.6
wt LII-11	4	-7.8 ± 0.3
LV-2 LII-11	2	-10.6 ± 0.2
wt DI-3	3	-16 ± 0.4
LV-2 DI-3	1	-18.9 ± 0.0
wt DI-8	5	-9.6 ± 0.4
LV-2 DI-8	2	-12.5 ± 0.1
LV-2 max 1	8	-9.0 ± 0.4
LV-2 max 2	2	-10.7 ± 0,.5
LV-2 max 3	2	-10.5 ± 0.7
LV-2 max 4	2	-10.7 ± 0.3
wt max 5	2	-7.2 ± 0.6
wt max 6	7	-7.6 ± 0.4
wt max 7	3	-6.0 ± 0.5
wt max 8	3	-6.2 ± 0.5

Table 8-13: The number of predicted secondary structures for the short 5'-UTR sequences given by the Quickfold software, as well as the average free energy (ΔG) including standard deviation.

Strain	Number of suggested structures	Average ΔG [kcal/mol] + standard deviation
pIB11	2	-1.9 ± 0.3
LV-2	1	-4.6 ± 0.0
Tn25	1	-4.1 ± 0.0
Tn47	2	-2.1 ± 0.1
Tn58	2	-0.9 ± 0.2
LII-11	2	-1.7 ± 0.6
DI-3	1	-10.6 ± 0.0
DI-8	1	-6.2 ± 0.0

J. The putative promoter element present in some of the 5'-UTR fusions

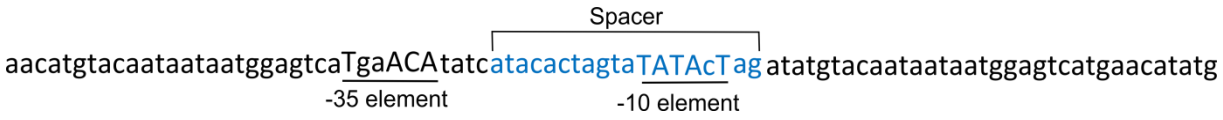

Figure 8-6: The putative promoter element present in some 5'-UTR fusions. The nucleotides corresponding to the consensus of the -10 and -35 region is in uppercase

Part Two: The Literature-Based Study

The Role of the Nucleotide Sequences at the 5'-ends of Genes in the Control of the Level of Recombinant Expression in *Escherichia coli*

Abstract: The nucleotide sequence at the gene 5'-end has a great influence on the expression level of genes, being the location of central mechanisms like transcription and translation initiation. Because of this the 5'-end sequence is an important target when designing genes for recombinant expression. This review will focus on recent research trends, covering the traits of the 5'-end that influence gene expression, as well as on approaches and tools targeting this region that have been utilized or show potential to be used to achieve desired recombinant expression levels in *E. coli*. In recent years it have become evident that the entire 5'-untranslated region as well as the initial coding sequence has great influence on gene expression, showing that there is more to designing genes for recombinant expression then picking a strong promoter and an optimal SD sequence.

Sammendrag: Nukleotidsekvensen til 5'-enden av gener har en stor innflytelse på genuttrykket, siden det er lokasjonen til sentrale mekanismer som transkripsjons og translasjons start. Dette gjør sekvensen til 5'-enden til et viktig mål ved design av gener for rekombinant produksjon. Denne artikkelen vil fokusere på nyere trender innen forskning som omfatter egenskaper ved 5'-enden som påvirker genuttrykket, samt tilnærminger og programvare rettet mot 5'-enden som har blitt brukt eller viser potensial til å bli brukt for å oppnå ønsket rekombinant genuttrykk i *E. coli*. I senere år har det blitt tydelig at hele den ikke-kodende 5'-enden samt den innledende kodende sekvensen har stor innflytelse på genuttrykket, som viser at det er mer til design av gener for recombinant uttrykk enn å kombinere en sterk promoter med en optimalisert SD-sekvens.

1 Introduction

E. coli has since the late 70s been extensively used for recombinant expression of proteins. It is an attractive production host due to its well characterized genome, well developed cloning tools and expression systems as well as industrial advantages such as rapid growth, cheap medium requirements and potential for high yield of the desired product. Close to 30% of bio-pharmaceuticals are produced in *E. coli* (Ferrer-Miralles et al. 2009) showing that it has maintained its important role as a recombinant microorganism.

Several approaches are used to achieve the desired expression of recombinant genes in *E. coli*. However, regulation of gene expression is immensely complicated and can occur at several levels on the path from DNA to a functional protein as illustrated in figure 1, and it should be noted that only part of the big picture will be addressed in this review (Fig. 1, highlighted).

The 5'-end of a gene at the DNA and messenger RNA (mRNA) level plays an important part in determining the efficiency of its gene expression. The 5'-end (including part of the 5'-flanking region) includes the promoter element, the 5'-untranslated region (5'-UTR) stretching from transcription start to translation start, and the initial coding sequence. Being the site of transcription and translation initiation there is no surprise that many regulatory mechanisms exert its role at the 5'-end. As will be described in this review, complex regulatory systems can operate at the 5'-end and small changes in the sequence may have a big impact on gene expression. Because of these traits the 5'- sequence is an important target in altering the expression of recombinant genes. This review will focus on recent research trends covering the traits of the 5'-end region influencing gene expression, and also on approaches targeting this region that have been utilized or show potential to be used to achieve desired expression levels in *E. coli*. The focus on recent research means that well known regulatory elements such as the promoter and the ribosome binding Shine Dalgarno (SD) sequence are not the main focus of this review. Instead the importance of the remaining part of the 5'-UTR and the initial coding region and also the subject of RNA regulation will be brought to attention. Aspects influencing gene expression at the protein level are beyond the scope of this review, such as protein-protein interaction and protein folding even if the N-terminus is involved.

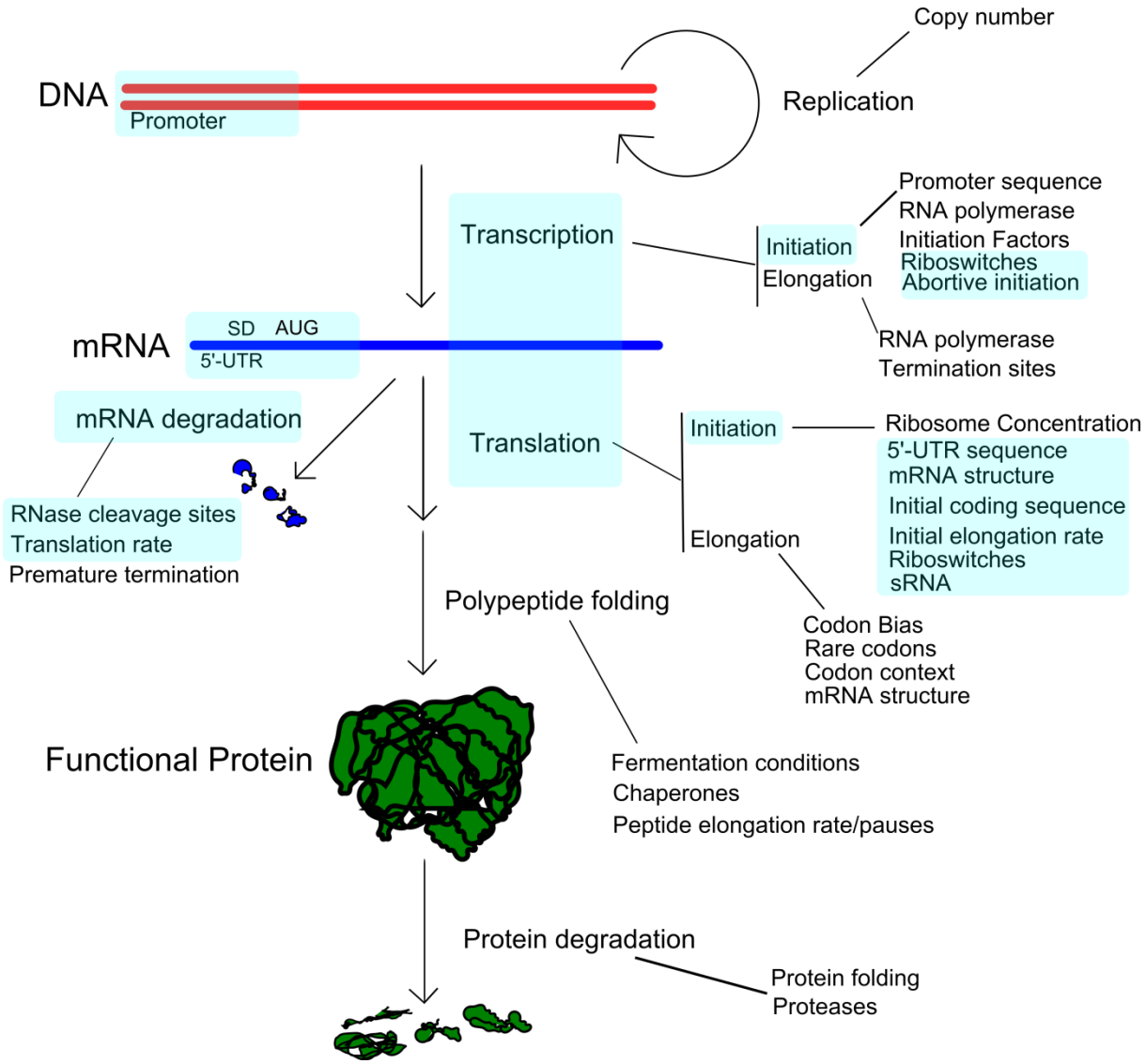

Figure 7: An overview of the synthetic pathway from DNA to a functional protein and some of the factors influencing gene expression. The focus areas of this review article are highlighted.

Since synthesizing synthetic gene sequences have become progressively easier over the years it has become commonplace to design synthetic genes, and other relevant DNA elements, when attempting to reach a specific expression level of a recombinant gene. However, a drawback in much of the research done on altering gene expression is that it is highly based on trial and error and only the success stories get published, often just presenting the wild type and the altered sequence. This makes it difficult to pin point the best way to design synthetic genes. Recently, studies analyzing large datasets of altered sequences with respect to their influence on gene expression have been published, some of which will be presented in

this review. Finally, the available bioinformatics tools aimed to design synthetic genes and commercial expression systems will be addressed.

2 The transcriptional level

The first step in protein synthesis is transcription where RNA polymerase (RNAP) initiates synthesis of messenger RNA (mRNA). The RNAP consists of five subunits of which the σ-factor is responsible for specific promoter recognition. *E. coli* produces several different σ-factors (mostly σ^{70}) recognizing different promoter sequences. The promoter can be divided into four regions: the two hexameres named the -10 and -35 region after their location in relation to the nucleotide first transcribed by RNAP termed the transcription start site (TSS), the spacer region between them (consensus length of 17 bp) and the A/T rich UP element (at position -40 to -60). The -10 and -35 regions are especially important for recognition of the σ-factor and the closer they are to the consensus sequence the stronger the promoter is (DeHaseth et al. 1998). The TSS chosen by the RNAP is dependent on the distance from the -10 region, but the sequence will also influence the start site selection, for instance purines are preferred over pyrimidines (Walker et al. 2002). A part of the initiation process is formation of the "open complex" where the two DNA strands melt (separate) from bp -11 to about +2/+3, but up to base +20 may be involved before RNAP escapes the promoter and forms the elongation complex (Hsu 2002; Davis et al. 2007). For more details on RNAP-promoter interaction see (DeHaseth et al. 1998). The importance of the promoter in recombinant expression is highly recognized. The most common commercially expression systems for *E. coli* utilize some version of the $P_{lac,}$ $P_{trp,}$ $P_{tac,}$ $\lambda P_{L,}$ λP_R and T7 promoter (see (Schumann et al. 2004) for details). Other factors recently found to influencing the transcription efficiency other than the basic interaction between the RNAP and the promoter sequence are addressed below.

The effect of several promoter elements in control of a gene has been investigated by incorporating clusters of core-tac-promoters in tandem. The transcript amount had a stepwise increase correlating with the number of promoter elements, until the maximum enhancement was reached at five promoters (Li et al. 2012).

The first nucleotides of the mRNA transcribed makes up the 5'-UTR, which for long has been known to influence translation. However, it has been discovered through recent studies that even mutating the 5'-UTR sequence more than 7 bp away from the TSS can result in a significant increase in transcript amount without changing the mRNA stability or the TSS position. An increase in the transcript amount of the penicillin resistance gene *bla* of up to 16-fold compared to wild type transcript amounts has been reported. It is speculated on that this effect may come from a faster formation of the open transcription initiation complex and/or faster promoter escape by RNAP (Berg et al. 2009).

Failure of promoter escape is linked with a mechanism termed "abortive initiation", which is the cyclic premature release of short (2-15 bp) RNA sequences by the RNAP. The phenomenon has long been observed *in vitro*, but was quite recently proven to take place *in vivo* as well (Goldman et al. 2009). A promoter close to the consensus sequence will lead to more abortive initiation due to tight binding of the RNAP preventing promoter escape. Hence, the promoter sequence is the primary regulator of abortive initiation, but in addition, the initial transcribed region can serve as a secondary regulator by altering the escape pattern if transcription is limited by promoter escape (Hsu et al. 2006). As well as influencing the transcription efficiency directly, abortive initiation can have an additional effect on gene expression posed by the aborted transcripts produced. It is shown that the aborted transcripts may regulate gene expression through trans-acting antitermination activity. For instance the aborted transcripts produced from the T7 promoter interact with the T7 terminator Tφ, disrupting its termination of transcription, which allows for the expression of genes downstream from it (Lee et al. 2010). A class of Short RNA oligonucleotides of 2-5 bp, termed "nanoRNA", to distinguish them from other classes of small RNAs (Mechold et al. 2007), can be the product of abortive initiation. They are shown to influence gene expression through priming of transcription initiation and altering of the sequence and phosphorylation state of the 5'-end. However, the extent of these regulatory mechanisms is uncertain. Sources of nanoRNAs other than from abortive initiation can be intermediate products from RNA degradation (Nickels et al. 2011).

3 Degradation of mRNA

mRNA is readily degraded in the cellular environment; hence the rate of mRNA decay together with transcription efficiency will determine the transcript amount available for translation. mRNA degradation involves several components forming the degradosome, where the most important component is RNase E catalyzing the rate limiting step of cleaving single stranded A/U rich regions. The cut regions leave fragments with monophosphorylated 5'-ends which are more easily degraded than triphosphate ends (Arraiano et al. 2010). It is important to realize that the steps in protein synthesis influence each other as is the case when ribosomes covering an mRNA during translation protect it from degradation. Hence, mechanisms decreasing or blocking translation might increase mRNA decay (Yarchuk et al. 1991).

Through a study of the *E. coli* transcriptome Lenz and colleagues described a correlation between low GC-content and less stable secondary structures in the 5'-mRNA with high mRNA half life. The mechanism at play might be a protective effect as a result of RNA binding proteins known to bind A/U rich regions (Lenz et al. 2011).

RNA transcripts can serve a role as regulators in bacteria. These non-coding regulatory transcripts are relatively short (50-300 bp) and denoted as small RNAs (sRNAs). Most is known of the group of sRNAs called trans-encoded sRNAs or antisense RNA which regulate by base-pairing with mRNA. Most of these bind at the mRNA 5'-end at or close to the ribosomal binding site (RBS). For details on sRNA, see the review by Stortz et al. 2011 (Storz et al. 2011). sRNA blocking translation will increase mRNA decay due to less protection provided by the ribosomes covering it. However, there are examples of sRNA increasing mRNA decay as part of an active mechanism as opposed to a secondary effect from blocked translation. The RNA chaperone protein Hfq is known to bind RNase E at the C-terminal end as well as certain sRNAs targeting the mRNA 5'-end such as RyhB involved in rapid degeneration of target mRNAs. It was discovered through a study involving RyhB that the increase in decay of the target mRNA after binding of RyhB at the RBS region was greater than what could be explained by blocked translation alone, revealing an active mechanism directed towards the 5'-end of target mRNAs increasing their decay rate (Prévost et al. 2011).

4 The translational level

The direct synthesis of a protein peptide chain through assembly of amino acids brought to the ribosomes by tRNAs occurs in the translation step. The rate limiting part of translation is translation initiation taking place at the mRNA 5'-end (Simonetti et al. 2009). The central mRNA elements are the Shine Dalgarno (SD) sequence, with consensus sequence GGAGG (in *E. coli)*, and the start codon where the most common and effective codon is AUG (Ma et al. 2002; Laursen et al. 2005). The small ribosomal subunit (30S) binds the mRNA through base pairing between the SD sequence and the 3'-end of the 16S ribosomal RNA (rRNA). The initiation factors (IF1, IF2 and IF3) and the initiator tRNA (fMet-tRNA), binding the start codon, joins the 30S and in combination with adaptation to the mRNA 5'-end results in the formation of the 30S initiation complex. Then the large ribosomal subunit (50S) is recruited, followed by release of the initiation factors and synthesis of the peptide chain (Simonetti et al. 2009). For more details on the mechanisms of translation initiation the following reviews can be recommended (Simonetti et al. 2009; Malys et al. 2011).

Even though translation and transcription often are focused upon separately it is important to note that they are indeed coupled mechanisms in bacteria. It is evident that the moving ribosome increases the transcription rate by preventing spontaneous backtracking by RNAP when the ribosome is in proximity of the RNAP. Thus, mechanisms slowing down the ribosome might decrease the transcript amount as well as translation rate (Proshkin et al. 2010). Not much is known of the mechanistic details of the crosstalk between ribosome and RNAP, but tools have recently been developed for exploring this interaction (Castro-Roa et al. 2012).

4.1 The 5'-UTR sequence and the necessity of the Shine Dalgarno

Much attention has been given to the undoubtedly important SD sequence, but recent research has emphasized the importance of the rest of the 5'-UTR as well. β-lactamase activity has been increased more than 20-fold by altering the 5'-UTR sequence. The alterations did not change the putative SD sequence and no correlation with secondary structures introduced or removed were found. A majority of the mutations were A to C and T to C, contradicting the common assumption of an A and T/U preference in the 5'-UTR due to reduced formation and

strength of secondary structures (Berg et al. 2009). The same strategy has successfully been applied to other genes as well. The effect is highly gene-dependent but a small set of 5'-UTR variants seems to work for most genes (Berg et al. 2012). Maximizing the gene expression is not always desired in recombinant production, due to factors like toxicity, the metabolic burden resulting from over-expression and the necessity of fine tuning expression in metabolic engineering. For the latter, targeting the 5'-UTR has proven to be a useful tool since 5'-UTR variants have been shown to downregulate induced gene expression without losing the inducible qualities, and to downregulate constitutive promoters as well as controlling the metabolic flow of the carotenoid sarcinaxanthin production pathway (Lale et al. 2011).

As more prokaryotic genomes have been sequenced it has been discovered that 5'-UTRs without a SD sequence are common in prokaryotes, suggesting more prokaryotic translation initiation mechanisms than previously thought (Chang et al. 2006). In a minimal, reconstituted *E. coli* system the majority of the effective RBSs were C-rich non-SD sequences showing high complementarity with the G-rich 16S rRNA. The C-rich non-SD RBS sequences tested in this study did not yield high level gene expression *in vivo* (Barendt et al. 2012). However, there are examples of higher recombinant expression levels achieved from a non-SD RBS sequence compared to the conventional SD sequence in *E. coli* (Mironova et al. 1999).

4.2 Stable secondary structures at the 5-end reduce gene expression

The research of Kudla and colleagues (Kudla et al. 2009) points to the mRNA stability of the 5'-end as a key feature affecting the translation initiation rate. They found a correlation between less stable mRNA structures and high translation initiation rate. The study involved the creation of a synthetic library consisting of 154 versions of the green fluorescent protein (*gfp*) gene only differentiating by random synonymous mutations. The mRNA folding energy was predicted using a moving window analysis. The analysis identified the region from nucleotide -4 to +37 relative to the translation start site which explained close to 10-times as much of the 250-fold variation in protein amount as any other predictive variable. To further emphasize the effect of the mRNA stability, low expressive *gfp*-variants gave higher expression when fused with a 28-codon tag with low mRNA stability (Kudla et al. 2009). The dataset obtained by Kudla et al. was later reanalyzed resulting in a model taking both 5'-

mRNA stability and codon bias into account, which better explained the variance in expression of the data set. It was concluded that codon usage was relevant in sequences without a strong 5'-mRNA structure (Supek et al. 2010). In a large study looking at the complete genomes of 340 species, a universal trend of reduced mRNA stability of the first 30 to 40 nucleotides of the coding region in all cellular life was found. The difference was largest in genes with high GC content and for the genes of prokaryotes with low optimum growth temperature. This makes thermodynamically sense since high GC content and low temperatures induce stable mRNA structures (Gu et al. 2010). It is hypothesized that the regulatory effect on gene expression mainly originates from interference with the start codon rather than ribosome binding to SD since the region of reduced mRNA stability does not overlap with the SD sequence (Kudla et al. 2009; Gu et al. 2010). The expression level of several wild type proteins has been significantly increased in *E. coli* by reducing the 5'-mRNA stability through the introduction of synonymous mutations in the first 72 nucleotides of the coding region. This resulted in development of a model predicting the suitability of expression of recombinant genes in *E. coli*, suggesting engineering of the 5'-end if the free energy (ΔG) is more negative than -17 to -20 kcal/mol (Cèbe et al. 2006). Thermostable enzymes originating from thermophilic microbes are of industrial interest, but because of cultivating difficulties it is advantage to express these genes in alternative hosts such as *E. coli.* Reducing the mRNA stability of the 5'-mRNA in combination with increased temperature have been utilized to increase the recombinant expression level of a thermostable enzyme originating from the thermophilic microbe *Deinococcus geothermalis* in *E. coli* (Szeker et al. 2010). The authors suggest this as a strategy for overexpressing genes limited by translation inhibiting 5'-mRNA structures and emphasizes that it can be especially useful for recombinant expression of genes originating from thermophiles, due to less selection pressure for unstable 5'-mRNA structures in their native high temperature environment. The strategy of reducing the 5'-end stability has been used to achieve efficient expression of the human aglycosylated FcγR1 Receptor in *E. coli.* The synthetic codon-optimized FcγR1 gene did not reach the same high expression level of the other recombinant FcγR genes in the study. However, the problem was solved by reducing the ΔG of the 100 nucleotides downstream of the promoter, including the first 21 codons, from -22.5 to -12.4 kcal/mol (Jung et al. 2010). The recombinant expression level of the hormone somatotropin has been greatly enhanced by modification of the +2 to+8 codon region by incorporating synonymous mutations. Expression levels of up to 23% of total cell protein (TCP), measured by coommasie blue staining, were achieved by introducing A or U at wobble positions when

possible. A correlation with high expression levels and low structural energy could be found for most but not all of the 21 gene variants (Sadaf et al. 2008). High recombinant expression levels of the human *G-CSF* gene in *E. coli* were achieved by optimizing the AT-content of the ten first codons. In all, 128 sequences were analyzed and the expression level correlated with the AT-content to an optimum expression reached at 51.8%, resulting in 14-fold greater gene expression (protein amount) than the native sequence. Considerable variation was observed between genes with different codons and approximately the same AT-content, and codon number 2,3,4,5 and 10 were found to be responsible for most of the positive effect. Codon optimizing the same region using a codon usage table gave no positive effect on gene expression (Krishna Rao et al. 2008). Analysis of regional tendencies in the coding region of genes derived from 816 bacterial genomes discovered a bias towards high AT-content, low amount of mRNA secondary structure and low CAI in the 5'-end of the coding region (first 35 bases). CAI is short for "Codon Adaptation Index" which is based on giving the most common codons in the host organism genome a favourable score (Sharp et al. 1987). Optimizing the CAI is a common tactic for codon optimizing a recombinant gene. The analysis was followed up by the construction of 285 synthetic sequences coding for three different test proteins designed to explore the region specific influences on the genetic expression. The gene expression was highly dependent on high AT-content and low mRNA secondary structure content at the 5'-end of the coding region and a high CAI value was found to contribute to a lesser extent. It is speculated on that the mechanism involved might be the recruitment of RNA helicases recognizing AU rich regions (Allert et al. 2010). An expression stimulating fusion tag termed an "expressivity tag" has been made from a 21 nucleotide long segment of the *infB* gene (coding for IF2). The fusion tag enhanced the production of three different proteins when fused to the 5'-end, thereof two proteins of industrial interest which are difficult to express in *E. coli.* The fusion led to a significant reduction in the 5'-mRNA stability which is hypothesized to be the cause of the positive effect on expression level (Hansted et al. 2011).

Studies indicate that translation initiation depends on certain segments of the mRNA to be single stranded for successful base pairing of the SD sequence and start codon to the 16S rRNA 3'-end and initiator tRNA anti codon, respectively. However, there are examples of efficiently translated genes despite stable secondary structures at the SD and start codon. One possible solution for this paradox is the "ribosome standby model" assuming that the 30S ribosomal subunit is able to bind single stranded mRNA regions flanking the secondary

structures sequestering the ribosome binding regions. The single stranded mRNA region can then function like a loading site when stable mRNA structures eventually loosen (De Smit et al. 2003). Since the model was presented in 2003 more experimental data supporting the model have been published (Unoson et al. 2007; Loh et al. 2012).

In recent years attention has been brought to the existents of guanine rich four stranded structures called G-quadruplexes in the 5'-UTR, as opposed to the more extensively studied double stranded RNA elements (Bugaut et al. 2012). Most of the research has been directed towards human cells and other eukaryotes, but experiments introducing quadruplexes in the 5'-UTR of *gfp* in *E. coli* resulted in heavy reduction of gene expression (fluorescence). The effect could be lightened by reducing the strength of the structure. Another effect was that the expression also became more temperature sensitive (Wieland et al. 2007).

4.3 The importance of the initial coding sequence

Resent research indicate that the initial coding sequence has a greater influence on gene expression than the rest of the coding sequence. Attention has been directed towards possible interaction between the 16S rRNA and the initial coding region of the mRNA 5'-end. This is due to the identification of a sequence stretch in several *E. coli* genes termed the "downstream box" (DB) showing complementarity to the nucleotides 1469-1483 of the 16S rRNA named the "anti downstream box" (ADB) (Sprengart et al. 1996). Researchers have shown that an optimized DB can enhance protein synthesis of luciferase by 13-fold compared to a non-optimized DB in a construct without a functional SD sequence (Rush et al. 2005). An optimized DB also has a stimulatory effect in concert with SD, with a reported β-galctosidase activity increase of 2-3 fold (Etchegaray et al. 1999). An expression vector utilizing the strong T7 promoter has been modified with an optimized DB sequence as part of the vector. This resulted in an increase of recombinant gene expression by 35-70% for 4 different genes (Zhang et al. 2003). In other uses; constructs with a common DB have successfully been utilized to achieve co-expression of genes in *E. coli* extract where otherwise one gene product would dominate (Keum et al. 2006). Even though several examples of the DB effect exist the mechanism of the DB enhancement is uncertain, and evidence against the occurrence of base pairing between DB and ADB during translation initiation have been put forward (O'Connor et al. 1999; Moll et al. 2001). Regardless of the mechanism it seems like the early coding region has some effect on gene expression exceeding the rest of the coding sequence. This is

supported by the identification of a marked non-randomness in the 15-17 nucleotides downstream from the start codon in the *E. coli* genome, and analysis of the results support that the effect is based on the nucleotide sequence rather than codon bias (Fuglsang 2004). Further demonstrating the effect of the early coding region on gene expression have been performed by randomizing the sequence of the +2 and +3 codon position of four recombinant genes, achieving low, medium, and high gene expression with up to a 70-fold variation in expression. The effect of the randomized sequences on gene expression appeared to be gene specific (Ahn et al. 2008).

4.4 Codon usage in the initial coding sequence

In resemblance to the research done by Kudla and colleagues (section 4.2) (Kudla et al. 2009), Welch et al have studied factors affecting gene expression by making gene libraries only differentiating by synonymous mutations (Welch et al. 2009). 40 versions of two different genes were made resulting in up to a 40-fold variation in gene expression (protein amount). In this study the main contributor to the variation was found to be codon usage. However, the common tactic of optimizing the CAI did not correlate with high gene expression, but rather using the codons read by the tRNAs which are most highly charged during shortage of amino acids. The authors do not disprove the findings of Kudla and colleagues since the genes in this study had low mRNA stability at the 5'-end to begin with and most of the variants had significantly weaker stability than what was found to have an effect on *gfp* expression. In addition, some of the sequences with lower expression than expected by the codon based predictions had stronger than average mRNA stability near the initiation site and the predicted expression levels were achieved by replacing the fist 15 codons with codons generating less stable structures (Welch et al. 2009). The influence on gene expression resulting from codon bias and 5'-mRNA folding energy has been further addressed, and the conclusion was that the translation efficiency is determined by both factors. However, for endogenous *E. coli* genes codon bias was found to be the main regulatory factor due to a global selection for weak structural energy at the mRNA 5'-end (Tuller et al. 2010b). Research analyzing the codon usage, using the tRNA adaptation index (dos Reis et al. 2003), has found a universal trend of low efficiency codons in the 30 to 50 first codons of the mRNA 5'-end (except for the second codon). This feature is termed "the low efficiency ramp" (or just "ramp" for short) and the mean ramp length for prokaryotes is 24 codons. Since less effective codons will slow the ribosomes down, it is hypothesized that the ramp hinders collision of the ribosomes (ribosomal jamming). Such ribosomal collisions can lead to abortion of translation, so

reducing that phenomenon will make for a more efficient translation and lower the metabolic burden on the cell. Thus, making it a relevant factor for recombinant gene expression (Tuller et al. 2010a). Results from further research indicate that mRNA secondary structures and amino acid charge contributes to the ramp functionality. The ribosome is slowed down by positively charged amino acids, due to the negative electrostatic potential of the ribosomal exit tunnel, and from strong secondary structures, both more prominent in the ramp region except for weak mRNA folding at the first few codons (Tuller et al. 2011).

It is known that most proteins are translated with an AUG codon, coding a methionine (Met), as the fist codon. However, findings involving analyses of proteins in the Swiss-Prot databank suggest that there is a functional selection for the second codon as well. The selection is weaker than the one for the first codon but stronger than the selection of further downstream codons. The second codon preference is species dependent. Lysine (Lys) and serine (Ser) are the most common in prokaryotes with Lys at 16.7% and Ser at 13.6% in *E. coli*. However, variation within prokaryotic species was observed, some with Ser as the most abundant second codon amino acid. The species dependent nature of the second codon amino acid makes it a possible target in achieving successful recombinant gene expression (Shemesh et al. 2010). Recombinant expression of the hormone somatotropin has been greatly increased by modifying the initial codons. The +2 codon seemed to be an especially potent target, since a deletion of the +2 GCC codon moving TCC (Ser) into the + 2 position resulted in a 30% TCP expression. Nucleotide substitutions to ACC and CCC resulted in TCP of 27% and 28% respectively while UCC resulted in undetectable expression by Coommasie blue staining (same as wt). However, the largest expression increase was achieved by Introducing six histidine (His) codons in the place of the +2 codon built up from three CAUCAC repeats resulted in expression at 48% TCP (Sadaf et al. 2008). Another large survey including 442 genomes across three domains show less conservation of the second base rather than the first for the second codon. This is opposite of what is normally observed in codons suggesting a different selection pressure for the second codon compared to codons in general (Tang et al. 2010). A relationship between NGG codons (N = non G) in the +2 to +5 position and low expression of the *lacZ* reporter gene has been found. The effect is hypothesized to be a result of peptidyl tRNA drop-off triggered by NGG in early position. No reduction in expression was observed with NGG in position +7 and onwards (Gonzalez De Valdivia et al. 2005).

4.5 RNA thermometers

The temperature will affect all cellular processes including gene expression by influencing factors like mRNA stability, but some genes have such a fine tuned temperature dependent regulation that the term RNA thermometers (RNATs) has arisen. These genes involve virulence, heat shock and cold shock genes and all take advantage of temperature dependent structural changes entirely or partly involving the 5'-UTR. RNATs can be sensitive enough to detect a 1 ºC difference in temperature (Kortmann et al. 2012). There are several different classes of RNATs (reviewed elsewhere (Narberhaus 2010; Kortmann et al. 2012)), but all depend on structures making the SD sequence and/or initiation codon more or less available for translation. The mechanism does not solely depend on the favourability of low 5'-mRNA stability since lower temperature increasing the stability may lead to more favourable structures as seen in cold shock genes. RNATs are known to regulate gene expression in a gradually or zipper like manner, or as more of an on/off or switch like response giving them diverse possibilities as a biotechnological tool. Regulation of gene expression without the need of an chemical inducer may help to reduce cost at industrial scale and prevent possible toxic effects (Kortmann et al. 2012). Synthetic RNATs have been made by a rational design approach followed by *in vivo* screening. RNATs performing at the same level or better compared to natural RNATs was obtained after two rounds of selection screening (Waldminghaus et al. 2008). Replacing the original 5'-UTR with the 5'-UTR of the cold shock protein cspA has shown a potential of increasing recombinant gene expression from the T7 promoter at low temperature (15 ºC) by a factor of two (Kim et al. 2008).

4.6 Small RNA

sRNA, introduced in chapter 3, is known to target the 5'-UTR through base pairing which interfere with ribosome binding. Examples of regulation of gene expression by artificial antisense RNA do exist, showing that this mechanism can be utilized to decrease the expression of a gene of choice. 70% silencing of *lacZ* was achieved using sRNA targeting its 5'-UTR (Alessandra et al. 2008). In order du reduce mRNA decay RNaseE expression has been successfully reduced using sRNA hybridizing with the 5'-UTR of the RNaseE transcript. The stability of the sRNA is important for effective silencing and research indicate that the presence of a SD sequence in the antisense RNA increases stability (Stefan et al. 2010). Acknowledgment of the potential of utilizing RNA as a biological tool has resulted in the

development of techniques allowing efficient *in vivo* recombinant RNA production in *E. coli* (Ponchon et al. 2011).

5 Riboswitches

The term riboswitch was originally described as a RNA metabolite-sensing genetic switch without the need of protein interaction (Nahvi et al. 2002). Riboswitches are located in the mRNA 5'-UTR and the binding of a metabolite ligand to the riboswitch induces structural changes in the mRNA effecting gene expression. It consists of two regions: the aptamer, which is the highly conserved ligand binding domain, and the regulatory part referred to as the expression platform which is overlapping or in close proximity of the aptamer. The natural occurring bacterial riboswitches are involved in environmental adaptation through metabolite sensing. The most common regulatory mechanisms are transcription termination through the creation of a GC rich stem followed by a poly U tail (Rho independent terminator) and translational regulation by affecting the availability of the SD sequence and/or start codon to the ribosome. More complex structures like riboswitches stacked in tandem or two aptamers in one expression platform result in increased sensitivity of the expression regulation posed by the riboswitch (Bastet et al. 2011; Breaker 2011). A focus area in riboswitch research has been the use of riboswitches as a possible drug target. Riboswitches are suitable targets for antimicrobial targeting. This is because riboswitches recognize small molecules, most of them are unique for bacteria, and riboswitches are often involved in regulating genes essential for microbial survival (Deigan et al. 2011). One example is the use of the thiamine analogue pyrithiamine to interfere with the riboswitch downregulating the genes involved with biosynthesis and transport of thiamine (Sudarsan et al. 2005).

Methods have been developed to engineer aptamers with high affinity and specificity to the ligand of choice by selecting from large pools of randomized oligonucleotides. One example is the aptamer binding the cheap and well characterized molecule theophylline. These synthetic aptamers can be incorporated to create functional riboswitches. While most riboswitches function by reducing gene expression, some synthetically made riboswitches increase gene expression as a response to the ligand, making them interesting as inducible systems. (Topp et al. 2010; Wittmann et al. 2012). A theophylline sensing synthetic riboswitch, with low background expression of the *lacZ* gene, showing a 36-fold expression

increase in the presence of theophylline has been engineered using a high throughput enzymatic assay approach (Lynch et al. 2007). An improved riboswitch showing a 96-fold increase, while remaining the low background expression, was engineered by screening combined with alteration of the SD sequence. The screening consisted of two steps: first using fluorescence activated cell sorting (FACS) followed by enzymatic assays (Lynch et al. 2009). Synthetic riboswitches functioning at the transcription level by disrupting the termination stem in the presence of theophylline resulting in increased expression of the reporter gene (*gfp*) have been engineered by the use of FACS and fluorescent assays (Fowler et al. 2008). Complex regulation can be achieved by synthetic riboswitches, exemplified by the engineering of a tandem riboswitch functioning both at the transcriptional and translational level in a manner where gene expression is downregulated at high and low ligand concentrations, resulting in maximum expression at intermediate ligand concentration (Muranaka et al. 2010).

6 Bioinformatics tools and expression systems

A large number of bioinformatics tools exist today, performing tasks like identification of genes and gene elements, structure predictions, and exploring sequence similarities through alignments. However, not too many are made for designing synthetic genes with a desired gene expression, but some notable examples will be presented in this section.

The Ribosome Binding Site (RBS) Calculator uses a thermodynamic model calculating the Gibbs free energy of ribosome binding to predict the translation initiation rate (TIR). It can be used to predict the TIR of an existing RBS, using the reverse engineering mode. In addition it can be used to design a RBS sequence with a desired TIR for a specific gene in the forward engineering mode by combining the thermodynamic model with a stochastic optimization model. The model considers an mRNA stretch of 35 nucleotides upstream and downstream from the start codon and the predicted TIR is given on a proportional scale from 0.001 to 100 000+. The software allows for incorporation of restriction sites or other wanted sequence stretches in the designed RBS by preventing alteration of given nucleotides (Salis 2011). Experimental validation of the RBS calculator involving over 100 predictions in *E. coli* show an accuracy within a factor of 2.3 covering a range of 100 000-fold. The reuse of an RBS

sequence with another gene may decrease the TIR by 500-fold emphasizing the importance of including the initial coding region (Salis et al. 2009).

An alternative mathematical model also focusing on the translation initiation step has been developed. This model utilizes steady state kinetics on the steps involved in translation initiation including global and regional mRNA folding/unfolding and ribosome binding. The model was experimentally verified by synthesizing 22 RBS regions for the *luxR* gene with a broad spectre of predicted translation efficiencies. The results gave a correlation coefficient of R=0.87 between experimental measured gene expression in *E. coli* and predicted translation efficiency (Na et al. 2010b). This mathematical model as well as the effect on translation efficiency resulting from the spacer length between the SD sequence and start codon has been incorporated in the publicly available software RBSDesigner (Na et al. 2010a).

ExEnSo (Expression Enhancer Software) is a free computer tool aimed to enhance the expression of recombinant genes in *E. coli* by reducing the mRNA stability at the RBS region. The software allows alterations of certain nucleotides and creates an *in silico* library followed by calculation of the free energies in the -70/+96 regions and then picks out the sequence with the most favourable free energy. The output is a 5' primer which can be used to PCR amplify the gene of interest followed by sub-cloning into an expression vector. The approach resulted in significantly enhanced protein expression for 8 of the 10 proteins tested (Care et al. 2008)

GeneDesigner is a software designed for creating synthetic DNA elements optimizing the gene expression by addressing several factors covering the entire mRNA sequence. The optimizing is performed by switching codons by using a codon usage tables, and avoiding certain sequence elements and strong secondary structures. Acknowledging the important role of the 5'-end, the software allows for it to be treated separately filtering out NGG codons, repeats and secondary structures as well as choosing A/T nucleotides in wobble positions (Villalobos et al. 2006; Welch et al. 2011).

One aspect of optimizing gene expression which might be overlooked is that most expression vectors in use today consist of natural genetic parts and other elements such as multiple cloning sites (MCS) not systematically optimized for production (Gustafsson et al. 2012). Gene expression has been improved by redesigning the MCS of a commonly used expression vector in yeast, a method believed to be relevant for prokaryotic vectors as well (Crook et al. 2011). It is therefore important to pay attention to the newly developed promoters, cloning sites, selection tags etc made by companies such as New England Biolab and Promega, or

alternatively look to organizations like Addgene (www.addgene.org) and Biofab (www.biofab.org) devoted to catalogue and distribute vector elements.

7 Concluding remarks

Advancement in sequencing technology over the past years has resulted in several completely sequenced genomes available for analysis, discovering non-random tendencies in the 5'-end previously not recognized. In addition the advancement in artificial sequence synthesis has allowed for the creation and analysis of large datasets of several synthetically made versions of a gene to better understand the factors determining gene expression. The number of such publications is still low. Thus, the number of genes analyzed in that manner need to increase to reach reliable conclusions. In addition the conclusions available today do not correspond completely.

The research field of trans-acting regulatory RNAs and riboswitches is still relatively fresh, at least in the context of applying them as biotechnological tools. Their place as tools in research and industry will surely be established in the years to come.

As the knowledge of the factors influencing gene expression increase, bioinformatics tools can become more reliable and approach the effectiveness of screening large libraries. Some approaches need to be re-evaluated when considering recent findings, like how to best codon optimize a gene. It is also important to recognize that vector elements like MCS, although practical, are not negligible and may influence gene expression.

It is evident that there is not one definitive answer on the best strategy to synthesis the gene's 5'-end resulting in the desired expressive qualities. As shown in this review many approaches can be utilized even when limited to targeting the 5'-end and it is not as simple as picking a strong promoter and optimal SD sequence. Over the last years it has become evident that the entire 5'-UTR in addition to the SD sequence as well as the initial coding region, have great influence on gene expression, though much is still uncertain regarding the mechanisms.

References:

Ahn, J. H., J. W. Keum, et al. (2008). "High-throughput, combinatorial engineering of initial codons for tunable expression of recombinant proteins." Journal of Proteome Research **7**(5): 2107-2113.

Alessandra, S., T. Alessandro, et al. (2008). "Artificial antisense RNAs silence lacZ in *E. coli* by decreasing target mRNA concentration." Journal of Biochemistry and Molecular Biology **41**(8): 568-574.

Allert, M., J. C. Cox, et al. (2010). "Multifactorial Determinants of Protein Expression in Prokaryotic Open Reading Frames." Journal of Molecular Biology **402**(5): 905-918.

Arraiano, C. M., J. M. Andrade, et al. (2010). "The critical role of RNA processing and degradation in the control of gene expression." FEMS Microbiology Reviews **34**(5): 883-923.

Barendt, P. A., N. A. Shah, et al. (2012). "Broad-specificity mrna-rrna complementarity in efficient protein translation." PLoS Genet **8**(3): e1002598.

Bastet, L., A. Dubé, et al. (2011). "New insights into riboswitch regulation mechanisms." Molecular Microbiology **80**(5): 1148-1154.

Berg, L., V. Kucharova, et al. (2012). "Exploring the 5'-UTR DNA region as a target for optimizing recombinant gene expression from the strong and inducible Pm promoter in *Escherichia coli*." J Biotechnol **158**(4): 224-230.

Berg, L., R. Lale, et al. (2009). "The expression of recombinant genes in *Escherichia coli* can be strongly stimulated at the transcript production level by mutating the DNA-region corresponding to the 5'-untranslated part of mRNA." Microb Biotechnol **2**(3): 379-389.

Breaker, R. (2011). "Prospects for Riboswitch Discovery and Analysis." Molecular Cell **43**(6): 867-879.

Bugaut, A. and S. Balasubramanian (2012). "5′-UTR RNA G-quadruplexes: Translation regulation and targeting." Nucleic Acids Research **40**(11): 4727-4741.

Care, S., C. Bignon, et al. (2008). "The translation of recombinant proteins in *E. coli* can be improved by in silico generating and screening random libraries of a -70/+96 mRNA region with respect to the translation initiation codon." Nucleic Acids Research **36**(1): e6.

Castro-Roa, D. and N. Zenkin (2012). "In vitro experimental system for analysis of transcription-translation coupling." Nucleic Acids Research **40**(6): 1-12.

Cèbe, R. and M. Geiser (2006). "Rapid and easy thermodynamic optimization of the 5′-end of mRNA dramatically increases the level of wild type protein expression in *Escherichia coli*." Protein Expression and Purification **45**(2): 374-380.

Chang, B., S. Halgamuge, et al. (2006). "Analysis of SD sequences in completed microbial genomes: Non-SD-led genes are as common as SD-led genes." Gene **373**(1-2): 90-99.

Crook, N. C., E. S. Freeman, et al. (2011). "Re-engineering multicloning sites for function and convenience." Nucleic Acids Research **39**(14): e92.

Davis, C. A., C. A. Bingman, et al. (2007). "Real-time footprinting of DNA in the first kinetically significant intermediate in open complex formation by *Escherichia coli* RNA polymerase." Proceedings of the National Academy of Sciences of the United States of America **104**(19): 7833-7838.

De Smit, M. H. and J. Van Duin (2003). "Translational standby sites: How ribosomes may deal with the rapid folding kinetics of mRNA." Journal of Molecular Biology **331**(4): 737-743.

DeHaseth, P. L., M. L. Zupancic, et al. (1998). "RNA polymerase-promoter interactions: The comings and goings of RNA polymerase." J Bacteriol **180**(12): 3019-3025.

Deigan, K. E. and A. R. Ferré-D'Amaré (2011). "Riboswitches: Discovery of drugs that target bacterial gene-regulatory RNAs." Accounts of Chemical Research **44**(12): 1329-1338.

dos Reis, M., L. Wernisch, et al. (2003). "Unexpected correlations between gene expression and codon usage bias from microarray data for the whole *Escherichia coli* K-12 genome." Nucleic Acids Research **31**(23): 6976-6985.

Etchegaray, J. P. and M. Inouye (1999). "Translational enhancement by an element downstream of the initiation codon in *Escherichia coli*." Journal of Biological Chemistry **274**(15): 10079-10085.

Ferrer-Miralles, N., J. Domingo-Espín, et al. (2009). "Microbial factories for recombinant pharmaceuticals." Microbial Cell Factories **8**(17): 1-8.

Fowler, C. C., E. D. Brown, et al. (2008). "A FACS-based approach to engineering artificial riboswitches." ChemBioChem **9**(12): 1906-1911.

Fuglsang, A. (2004). "Nucleotides downstream of start codons show marked non-randomness in *Escherichia coli* but not in Bacillus subtilis." Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology **86**(2): 149-158.

Goldman, S. R., R. H. Ebright, et al. (2009). "Direct detection of abortive RNA transcripts in vivo." Science **324**(5929): 927-928.

Gonzalez De Valdivia, E. I. and L. A. Isaksson (2005). "Abortive translation caused by peptidyl-tRNA drop-off at NGG codons in the early coding region of mRNA." FEBS Journal **272**(20): 5306-5316.

Gu, W., T. Zhou, et al. (2010). "A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes." PLoS Computational Biology **6**(2): e1000664.

Gustafsson, C., J. Minshull, et al. (2012). "Engineering genes for predictable protein expression." Protein Expression and Purification **83**(1): 37-46.

Hansted, J. G., L. Pietikäinen, et al. (2011). "Expressivity tag: A novel tool for increased expression in *Escherichia coli*." J Biotechnol **155**(3): 275-283.

Hsu, L. M. (2002). "Promoter clearance and escape in prokaryotes." Biochimica et Biophysica Acta - Gene Structure and Expression **1577**(2): 191-207.

Hsu, L. M., I. M. Cobb, et al. (2006). "Initial transcribed sequence mutations specifically affect promoter escape properties." Biochemistry **45**(29): 8841-8854.

Jung, S. T., T. H. Kang, et al. (2010). "Efficient expression and purification of human aglycosylated Fcγ receptors in *Escherichia coli*." Biotechnology and Bioengineering **107**(1): 21-30.

Keum, J. W., J. H. Ahn, et al. (2006). "The presence of a common downstream box enables the simultaneous expression of multiple proteins in an *E. coli* extract." Biochem Biophys Res Commun **350**(3): 562-567.

Kim, S. H., M. A. Heo, et al. (2008). "The effect of the cspA 5′-Untranslated region on recombinant protein production at low temperature." Biotechnology and Bioprocess Engineering **13**(3): 366-371.

Kortmann, J. and F. Narberhaus (2012). "Bacterial RNA thermometers: Molecular zippers and switches." Nature Reviews Microbiology **10**(4): 255-265.

Krishna Rao, D. V., J. V. Rao, et al. (2008). "Optimization of the AT-content of codons immediately downstream of the initiation codon and evaluation of culture conditions for high-level expression of recombinant human G-CSF in *Escherichia coli*." Molecular Biotechnology **38**(3): 221-232.

Kudla, G., A. W. Murray, et al. (2009). "Coding-Sequence Determinants of Expression in *Escherichia coli*." Science **324**(5924): 255-258.

Lale, R., L. Berg, et al. (2011). "Continuous control of the flow in biochemical pathways through 5' untranslated region sequence modifications in mRNA expressed from the broad-host-range promoter Pm." Appl Environ Microbiol **77**(8): 2648-2655.

Laursen, B. S., H. P. Sørensen, et al. (2005). "Initiation of protein synthesis in bacteria." Microbiology and Molecular Biology Reviews **69**(1): 101-123.

Lee, S., H. M. Nguyen, et al. (2010). "Tiny abortive initiation transcripts exert antitermination activity on an RNA hairpin-dependent intrinsic terminator." Nucleic Acids Research **38**(18): 6045-6053.

Lenz, G., A. Doron-Faigenboim, et al. (2011). "Sequence features of *E. coli* mRNAs affect their degradation." PLoS One **6**(12): e28544.

Li, M., J. Wang, et al. (2012). "A strategy of gene overexpression based on tandem repetitive promoters in *Escherichia coli*." Microbial Cell Factories **11**(19): 1-10.

Loh, E., F. Memarpour, et al. (2012). "An unstructured 5′-coding region of the prfA mRNA is required for efficient translation." Nucleic Acids Research **40**(4): 1818-1827.

Lynch, S. A., S. K. Desai, et al. (2007). "A High-Throughput Screen for Synthetic Riboswitches Reveals Mechanistic Insights into Their Function." Chemistry and Biology **14**(2): 173-184.

Lynch, S. A. and J. P. Gallivan (2009). "A flow cytometry-based screen for synthetic riboswitches." Nucleic Acids Research **37**(1): 184-192.

Ma, J., A. Campbell, et al. (2002). "Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures." J Bacteriol **184**(20): 5733-5745.

Malys, N. and J. E. G. McCarthy (2011). "Translation initiation: Variations in the mechanism can be anticipated." Cellular and Molecular Life Sciences **68**(6): 991-1003.

Mechold, U., G. Fang, et al. (2007). "YtqI from *Bacillus subtilis* has both oligoribonuclease and pAp-phosphatase activity." Nucleic Acids Res **35**(13): 4552-4561.

Mironova, R. S., J. Xu, et al. (1999). "Efficiency of a novel non-Shine-Dalgarno and a Shine-Dalgarno consensus sequence to initiate translation in *Escherichia coli* of genes with different downstream box composition." Microbiological Research **154**(1): 35-41.

Moll, I., M. Huber, et al. (2001). "Evidence against an interaction between the mRNA downstream box and 16S rRNA in translation initiation." J Bacteriol **183**(11): 3499-3505.

Muranaka, N. and Y. Yokobayashi (2010). "A synthetic riboswitch with chemical band-pass response." Chemical Communications **46**(36): 6825-6827.

Na, D. and D. Lee (2010a). "RBSDesigner: Software for designing synthetic ribosome binding sites that yields a desired level of protein expression." Bioinformatics **26**(20): 2633-2634.

Na, D., S. Lee, et al. (2010b). "Mathematical modeling of translation initiation for the estimation of its efficiency to computationally design mRNA sequences with desired expression levels in prokaryotes." BMC Systems Biology **4**(71): 1-16.

Nahvi, A., N. Sudarsan, et al. (2002). "Genetic control by a metabolite binding mRNA." Chemistry and Biology **9**(9): 1043-1049.

Narberhaus, F. (2010). "Translational control of bacterial heat shock and virulence genes by temperature-sensing mRNAs." RNA Biology **7**(1): 84-89.

Nickels, B. E. and S. L. Dove (2011). "NanoRNAs: A class of small RNAs that can prime transcription initiation in bacteria." Journal of Molecular Biology **412**(5): 772-781.

O'Connor, M., T. Asai, et al. (1999). "Enhancement of translation by the downstream box does not involve base pairing of mRNA with the penultimate stem sequence of 16s rRNA." Proceedings of the National Academy of Sciences of the United States of America **96**(16): 8973-8978.

Ponchon, L. and F. Dardel (2011). "Large scale expression and purification of recombinant RNA in *Escherichia coli*." Methods **54**(2): 267-273.

Prévost, K., G. Desnoyers, et al. (2011). "Small RNA-induced mRNA degradation achieved through both translation block and activated cleavage." Genes and Development **25**(4): 385-396.

Proshkin, S., A. Rachid Rahmouni, et al. (2010). "Cooperation between translating ribosomes and RNA polymerase in transcription elongation." Science **328**(5977): 504-508.

Rush, G. J. and L. M. Steyn (2005). "Translation enhancement by optimized downstream box sequences in *Escherichia coli* and Mycobacterium smegmatis." Biotechnology Letters **27**(3): 173-179.

Sadaf, S., M. A. Khan, et al. (2008). "Expression enhancement of bubaline somatotropin in *E. coli* through gene modifications in the 5′-end coding region." J Biotechnol **135**(2): 134-139.

Salis, H. M. (2011). "The ribosome binding site calculator." Methods in Enzymology **498**: 19-42.

Salis, H. M., E. A. Mirsky, et al. (2009). "Automated design of synthetic ribosome binding sites to control protein expression." Nature Biotechnology **27**(10): 946-950.

Schumann, W. and L. C. S. Ferreira (2004). "Production of recombinant proteins in *Escherichia coli*." Genetics and Molecular Biology **27**(3): 442-453.

Sharp, P. M. and W. H. Li (1987). "The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications." Nucleic Acids Res **15**(3): 1281-1295.

Shemesh, R., A. Novik, et al. (2010). "Follow the Leader: Preference for Specific Amino Acids Directly Following the Initial Methionine in Proteins of Different Organisms." Genomics, Proteomics and Bioinformatics **8**(3): 180-189.

Simonetti, A., S. Marzi, et al. (2009). "A structural view of translation initiation in bacteria." Cellular and Molecular Life Sciences **66**(3): 423-436.

Sprengart, M. L., E. Fuchs, et al. (1996). "The downstream box: An efficient and independent translation initiation signal in *Escherichia coli*." EMBO Journal **15**(3): 665-674.

Stefan, A., F. Schwarz, et al. (2010). "Shine-Dalgarno sequence enhances the efficiency of *lacZ* repression by artificial anti-lac antisense RNAs in *Escherichia coli*." J Biosci Bioeng **110**(5): 523-528.

Storz, G., J. Vogel, et al. (2011). "Regulation by Small RNAs in Bacteria: Expanding Frontiers." Molecular Cell **43**(6): 880-891.

Sudarsan, N., S. Cohen-Chalamish, et al. (2005). "Thiamine pyrophosphate riboswitches are targets for the antimicrobial compound pyrithiamine." Chemistry and Biology **12**(12): 1325-1335.

Supek, F. and T. Šmuc (2010). "On relevance of codon usage to expression of synthetic and natural genes in *Escherichia coli*." Genetics **185**(3): 1129-1134.

Szeker, K., O. Niemitalo, et al. (2010). "High-temperature cultivation and 5' mRNA optimization are key factors for the efficient overexpression of thermostable Deinococcus geothermalis purine nucleoside phosphorylase in *Escherichia coli*." J Biotechnol **156**(4): 268-274.

Tang, S. L., B. C. H. Chang, et al. (2010). "Gene functionality's influence on the second codon: A large-scale survey of second codon composition in three domains." Genomics **96**(2): 92-101.

Topp, S. and J. P. Gallivan (2010). "Emerging applications of riboswitches in chemical biology." ACS Chemical Biology **5**(1): 139-148.

Tuller, T., A. Carmi, et al. (2010a). "An evolutionarily conserved mechanism for controlling the efficiency of protein translation." Cell **141**(2): 344-354.

Tuller, T., I. Veksler-Lublinsky, et al. (2011). "Composite effects of gene determinants on the translation speed and density of ribosomes." Genome Biol **12**(11): 1-18.

Tuller, T., Y. Y. Waldman, et al. (2010b). "Translation efficiency is determined by both codon bias and folding energy." Proceedings of the National Academy of Sciences of the United States of America **107**(8): 3645-3650.

Unoson, C. and E. G. H. Wagner (2007). "Dealing with stable structures at ribosome binding sites: Bacterial translation and ribosome standby." RNA Biology **4**(3): 113-117.

Villalobos, A., J. E. Ness, et al. (2006). "Gene Designer: a synthetic biology tool for constructing artificial DNA segments." BMC Bioinformatics **7**(285): 1-8.

Waldminghaus, T., J. Kortmann, et al. (2008). "Generation of synthetic RNA-based thermosensors." Biological Chemistry **389**(10): 1319-1326.

Walker, K. A. and R. Osuna (2002). "Factors affecting start site selection at the *Escherichia coli* fis promoter." J Bacteriol **184**(17): 4783-4791.

Welch, M., S. Govindarajan, et al. (2009). "Design parameters to control synthetic gene expression in *Eschorichia coli*." PLoS One **4**(9): 1-10.

Welch, M., A. Villalobos, et al. (2011). Designing genes for successful protein expression. **498:** 43-66.

Wieland, M. and J. S. Hartig (2007). "RNA Quadruplex-Based Modulation of Gene Expression." Chemistry and Biology **14**(7): 757-763.

Wittmann, A. and B. Suess (2012). "Engineered riboswitches: Expanding researchers' toolbox with synthetic RNA regulators." FEBS Letters **586**(15): 2076-2083.

Yarchuk, O., I. Iost, et al. (1991). "The relation between translation and mRNA degradation in the *lacZ* gene." Biochimie **73**(12): 1533-1541.

Zhang, X., P. Guo, et al. (2003). "A vector with the downstream box of the initiation codon can highly enhance protein expression in *Escherichia coli*." Biotechnology Letters **25**(10): 755-760.