Universidade de Aveiro Departamento de Química
Ano 2011

Joana Pinto Fernandes **_De novo_ gene synthesis**

Síntese de genes _de novo_

Universidade de Aveiro Departamento de Química
Ano 2011

Joana Pinto Fernandes **_De novo_ gene synthesis**

Síntese de genes _de novo_

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Biotecnologia Molecular, realizada sob a orientação científica do Doutor Manuel António da Silva Santos, Professor Associado do Departamento de Biologia da Universidade de Aveiro e co-orientação do Doutor Jörg Christian Frommlet, Investigador de pós-Doutoramento do Departamento de Biologia da Universidade de Aveiro.

Aos meus pais,

o júri

presidente

Prof. Doutor João Manuel da Costa e Araújo Pereira Coutinho
Professor associado do Departamento de Química da Universidade de Aveiro

Prof. Doutor Manuel António da Silva Santos
Professor associado do Departamento de Biologia da Universidade de Aveiro

Doutor Jörg Christian Frommlet
Investigador de pós-Doutoramento do Departamento de Biologia da Universidade de Aveiro

Doutora Gabriela Ribeiro de Moura
Investigadora auxiliar do Centro de Estudos do Ambiente e do Mar da Universidade de Aveiro

Doutor Rui Miguel Pinheiro Vitorino
Investigador associado do Departamento de Química da Universidade de Aveiro

Acknowledgements

To Jörg Christian Frommlet, for helping me in all the work made during this last year and specially, for the good friendship that was created between us;

To Professor Manuel Santos, for receiving me so well on his lab and for helping in every single moment of this project;

To all the team from the RNA Biology laboratory of University of Aveiro. All of them are amazing and very competent people that always helped me since my first day on the lab;

To the mass spectrometry team of the Chemistry Department that did its best to help me in the last stage of my project;

To the Bioinformatics Group of University of Aveiro, for the development of the gene optimization tools used in this work;

To University of Aveiro and particularly to Biology and Chemistry departments which provided all the material and conditions that I needed to construct my thesis;

To Professor João Coutinho for listening my expectations during this important stage of my life;

To all my dear BEST Aveiro members, amazing friends that contributed a lot in my development as a person during this year;

And finally, I really special thanks to my family, boyfriend and friends, for all their love, patience and happiness!

Key-words Translation, mRNA, gene optimisation, heterologous protein expression, codon usage, codon context, protein solubility

Abstract

Due to the degeneracy of the genetic code, an average protein of 300 amino acids can be encoded by the truly astronomical number of more than 10^{150} different codon combinations; more than the estimated number of atoms in the observable universe. However, the choice between synonymous codons in the mRNA coding sequence is not random, but follows rules and has important functions in translation accuracy, efficiency and co-translational protein folding. This additional layer of information that mRNA primary structures encode is found in all three domains of life and is modulated by evolutionary forces as mutation and selection. Particularly, codon usage and codon context are strongly biased features in mRNA primary structure of different species. In heterologous protein expression, the translation instructions contained in the coding sequence, need to be recognized by the heterologous host, to avoid the production of insoluble, non-functional proteins. Factors such as differences in codon usage, and codon context between native and heterologous host must be considered, as well as the presence of rare codon rich regions and mRNA secondary structures. Gene optimisation is often the solution to overcome these difficulties and to obtain a higher yield of functional, correctly folded heterologous protein.

The present work aimed at studying the effects of rare codons in a *Plasmodium falciparum* lysyl-tRNA synthetase gene (*Pf* LysRS) on protein solubility in *Escherichia coli* and whether protein solubility can be improved through codon harmonisation. Furthermore, the effect of other parameters such as codon context on translation accuracy and efficiency were analysed using the β-galactosidase gene as another model.

Palavras-chave

Tradução, mRNA, optimização de genes,expressão heterologa de proteinas, utilização de codões, contexto de codões, solubilidade proteica

Resumo

Devido ao facto do código genético ser degenerado, uma proteína composta por 300 aminoácidos, pode ser codificada por um valor verdadeiramente astronómico de mais de 10^{150} combinações de codões; mais do que o número estimado de átomos no Universo observável. Contudo, a escolha entre codões sinónimos na sequência de mRNA não é aleatória, mas pelo contrário, segue regras e apresenta funções importantes ao nível da precisão e eficiência de tradução, bem como no *folding* co-traducional de proteinas.

Esta camada adicional de informação que a estrutura primária do mRNA codifica é encontrada nos três domínios da vida e é modelada por forças como a mutação e a selecção. Em particular, a utilização de codões e o contexto de codões são características fortemente enviesadas na estrutura primária do mRNA de diferentes espécies. Na expressão heterologa de proteínas, as instruções traducionais contidas na sequência codificante necessitam de ser reconhecidas pelo hospedeiro heterologo de forma a evitar a produção de proteínas insolúveis, não funcionais. Factores como diferenças na utilização de codões e contexto de codões entre o hospedeiro nativo e heterologo devem ser consideradas, tal como a presença de regiões ricas em codões raros e a presença de estruturas secundárias de mRNA. A optimização de genes é frequentemente a solução encontrada para ultrapassar estas dificuldades e para obter um maior rendimento em proteínas heterologas solúveis e funcionais.

O presente trabalho tem por objectivo estudar os efeitos de codões raros no gene da Lisil-tRNA sintetase de *Plasmodium falciparum* (*Pf* LysRS) em termos de solubilidade proteica em *Escherichia coli*, bem como estudar como essa solubilidade pode ser melhorada através de harmonização de codões. O efeito de outros parâmetros, como o contexto de codões, na precisão e eficiência de tradução foram analisados num outro modelo, o gene de β-galactosidase.

List of contents

1 Introduction

1.1 Protein Synthesis

Protein synthesis is a complex process, in which amino acids are linked sequentially through peptide bonds to form a polypeptide chain. The resulting proteins are macromolecules with important functions in all cellular processes, such as structure, storage, movement, transport, signalling and the catalysis of biological reactions.

The information to build proteins is carried by genes in the form of the nucleotide sequence and governed by the rules of the genetic code. Messenger ribonucleic acids (mRNAs) are key molecules of life, as they establish the link between a gene and the cell's protein factories. During the first step of protein synthesis, called transcription, a RNA complementary copy of a gene is created [1]. To synthesize a protein, that mRNA is attached to the ribosome, a large multi-molecular complex that performs the second step of protein synthesis – translation. The entire pathway from gene to protein is tightly regulated and the involved processes are controlled by many factors, including DNA chemical and structural modifications as well as transcription-, post-transcription- and translation-regulation [2]. For instance, in prokaryotes, transcription is regulated by activators, repressors, and in some cases enhancers and in both eukaryotes and prokaryotes translation initiation can be regulated by mRNA secondary structures that expose or sequester the ribosomal binding site (RBS) [3]. Since this study is mainly concerned with heterologous protein expression in prokaryotes, hence forth, focus will be given to the prokaryotic processes.

1.1.1 Transcription

In this first step of protein synthesis, a DNA sequence is read by a RNA polymerase, which produces a complementary RNA strand. To initiate the prokaryotic transcription, the RNA polymerase first binds to several specificity σ-factors to form an holoenzyme and then recognizes specific DNA sequences in the promoter region of the gene (-10 and -35 regions). After this stage, the DNA is unwound and the holoenzyme reads the DNA strand, synthesizing a single-stranded RNA transcript of the gene (fig. 1). Downstream of the coding region, the RNA polymerase reads some DNA inverted sequences (approx. 40bp) that encode stem loops. These structures are able to reduce RNA polymerase affinity and lead to transcription termination. Alternatively, in some prokaryotes the termination can be caused by a small hexamer protein (rho) that recognizes termination signals contained in the DNA sequence and forces the RNA polymerase to dissociate. The production of mRNA template is a repeated process, i.e. multiple RNA copies of the DNA template are produced in the cell. In prokaryotes, transcription occurs in the cytoplasm along side with translation.

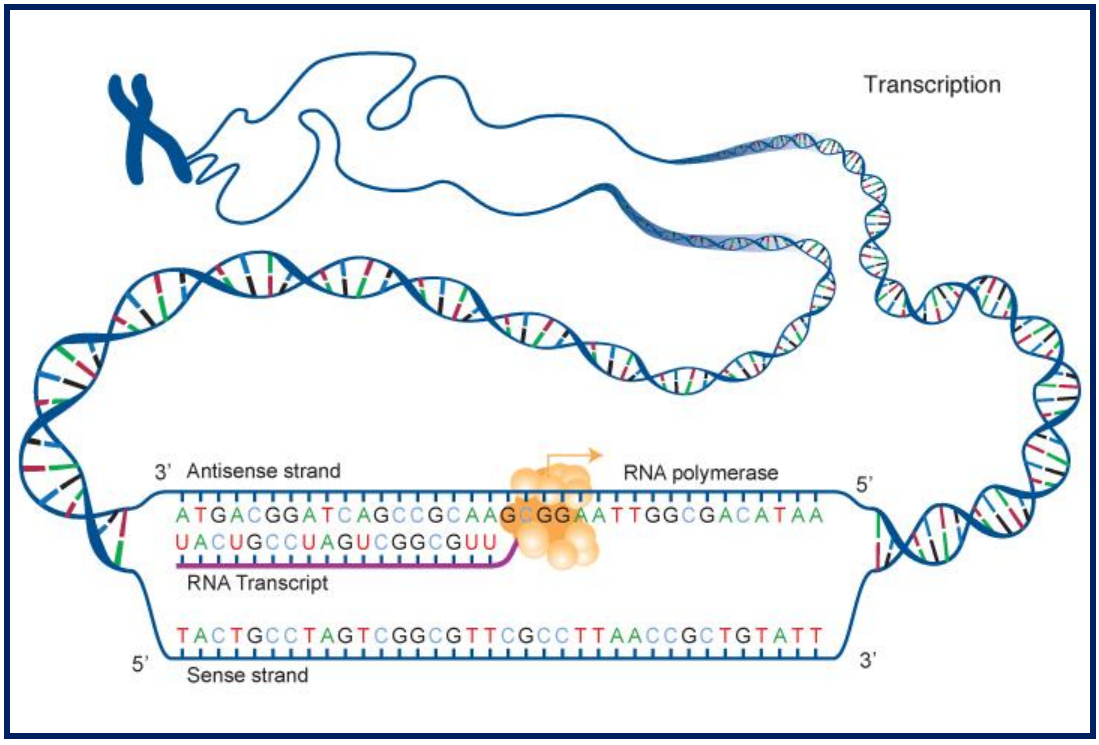


Figure 1. Basic scheme of the transcription process. (*National Human Genome Research Institute (NHGRI); www.genome.gov)*

1.1.2 Translation

Translation is the process through which the genetic information, carried by mRNA, is transferred into the amino acid sequence of a protein.

For this mediated production of the proteome [4], several components need to be recruited. According to the genetic code, each amino acid is encoded in sets of three bases, called ‘codons’ along the mRNA sequence. The message contained in the reading frame is then read by the ribosome, large molecular machinery where the proteins are produced (fig.2). Transfer RNAs (tRNAs) play an important role in translation as they are the adaptors that interact with mRNA on one end, and bind to amino acids on the other.

In general, protein biosynthesis is very similar in prokaryotes and eukaryotes as both systems use the same basic processes. However, some peculiarities exist, especially in the added complexity of the eukaryotic translation initiation system. While promoter specific initiation in eukaryotes requires several initiation factors, in bacteria, only a single polypeptide is required to bind the RNA polymerase.

Although it is generally accepted that protein synthesis is mainly controlled during transcription, the translation process is also regulated by a large range of factors. In particular, the translation initiation is mediated by three protein initiation factors (IF), designated IF-1, IF-2 and IF-3. However, on a different level of the process, other regulatory mechanisms are also important. For instance, excessive protein production and protein accumulation can shut down translation (auto-regulation) and depending on the environmental conditions and cell requirements, metabolic instability can cause mRNA to degrade rapidly.

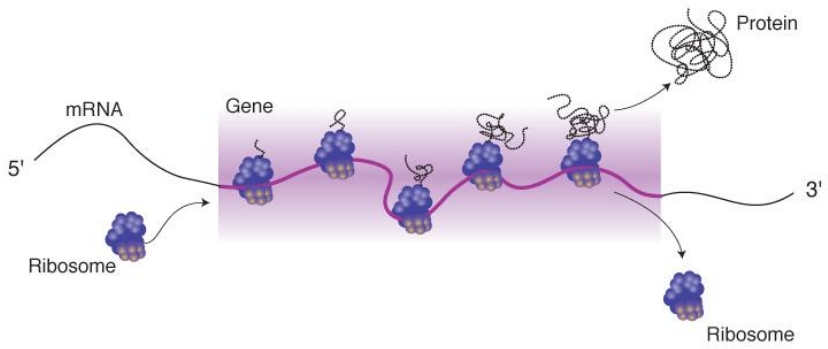


Figure 2. Representation of the translation process. (*NHGRI; www.genome.gov*)

The prokaryotic translation process can be divided in three main phases – **initiation**, **elongation** and **termination**:

- Initiation

The translation apparatus is a complex machinery, composed of multiple components, which have to be assembled before a functional unit is created. One of the key components of this machinery is the ribosome, composed of two ribosomal subunits, one large and one small: 50S and 30S. Additional components of the translation system are: the mRNA template, initiation factors and energy in form of GTP.

Upstream of the initiation codon, near the 5' UTR of the mRNA, lies an important sequence for translation initiation, called Shine Delgarno sequence (6-10 bases). This sequence, with the consensus ´AGGAGG´, is complementary to the 16S ribosomal RNA sequence ´CCUCCU´ and allows the correct binding of the 30S ribosomal subunit/initiation factor-3 complex to the mRNA. A short scanning in 3' end direction is performed until the small subunit finds the start codon. At the same time, an initiation factor-2 facilitates the attachment of the first tRNA to the start codon, which in prokaryotes is always N-formyl methionine (Met). Both initiation factors, IF-2 and IF-3 are stimulated by a third one, IF-1.

All of these components (small ribosomal subunit, initiator tRNA and IF-1/2/3) establish the initiation complex. At this stage, the large ribosomal unit joins the complex, a GTP molecule is hydrolysed and the initiation factors are released.

- Elongation

During the elongation of the polypeptide chain, the addition of amino acids occurs at the carboxyl end of the growing chain. Several ribosomes can read one mRNA molecule at a time, forming what is called a polysome. This means, that from a single mRNA, many polypeptides can be produced. The whole process requires the elongation factor EF-Tu (a small GTPase), and energy provided by GTP (3 GTPs per amino acid bond).

Three sites for tRNA binding are established in the ribosome, the peptidyl (P), aminoacyl (A) and exit (E) site. The first step of elongation is the binding of the initiating aminoacyl-tRNA to the 'P' site, with a conformational change that opens the 'A' site. This is followed by the binding of the complementary amino acid of the next codon to the 'A' site. The enzyme peptidyl transferase, contained in the large sub-unit of the ribosome, establishes a bond between the first and second amino acid. In the last stage of elongation, translocation, the ribosome moves 3 nucleotides (one codon's length) in the 3' end direction, bringing the newly formed peptidyl-tRNA to the 'P' site. The ribosome continues to translate the next codons as more aminoacyl-tRNAs bind to the A site, and before it reaches the stop codon of the sequence.

- Termination

Between the initiation codon (AUG) and the stop codon (UAG, UAA, and UGA) is the coding region of a gene, the open reading frame (ORF). When the stop codon is reached, release factors (namely, RF-1 and RF-2) read the triplet and trigger the hydrolysis of the ester bond in the peptidyl-tRNA. The complete polypeptide is released from the tRNA, the tRNA is released from the ribosome and the two ribosomal subunits separate from the mRNA. Finally, a third release factor (RF-3) forces the dissociation of RF-1 and RF-2.

1.2 mRNA primary and secondary structure

As mentioned earlier, protein expression is influenced by many factors and on all levels from transcription to protein folding (fig. 3). Some of these factors are directly encoded in the genetic message, i.e. the coding sequence itself plays an important role in gene expression dynamics. The present study focuses on these factors that are encoded within the mRNA primary structure and that affect protein synthesis.

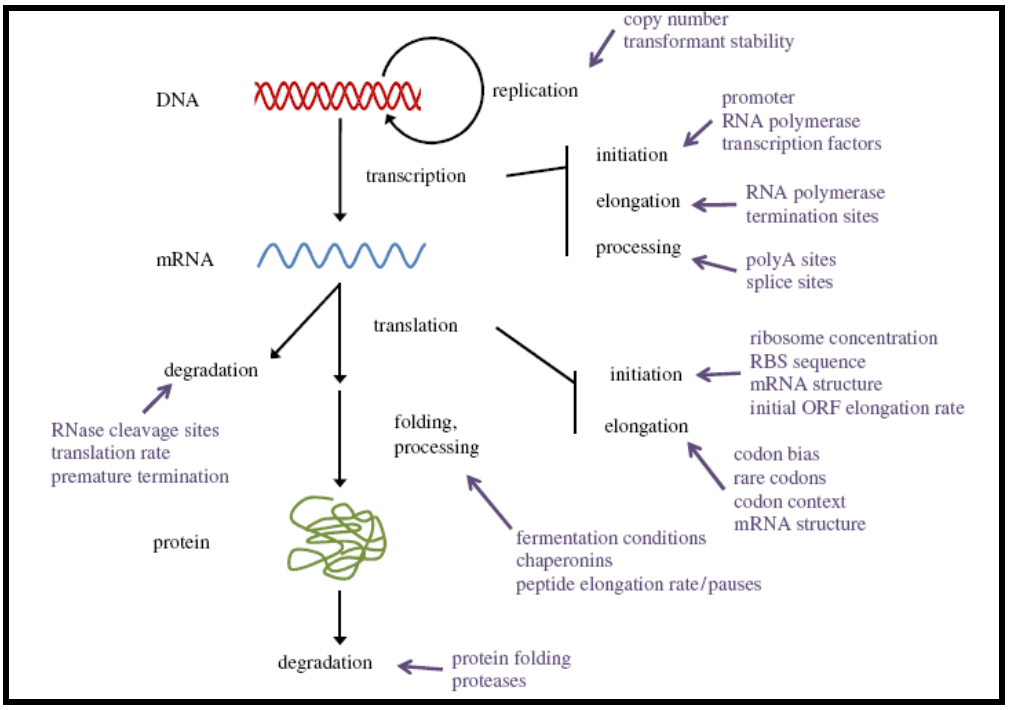


Figure 3. Representation of factors influencing protein expression. [2]

Particularly during the elongation step of translation, several features, such as **codon usage, codon context, codon correlation** and **mRNA secondary structure**, have been identified to clearly affect translation efficiency and accuracy [5-8]. Translational efficiency is related to the speed with which a message is translated and can affect protein yield but also the number of ribosomes that are available to translate other messages. Translation accuracy is the rate with which amino acids are correctly incorporated, according to the genetic code. High accuracy benefits the cell by reducing mistranslation products with potentially lowered functionality and the toxicity of harmful mistranslation products [5]. Because translational efficiency and accuracy are so fundamentally important parameters, features such as codon usage and codon context are crucial for the proper functioning of the cell.

1.2.1 Codon usage

With the exceptions of methionine and tryptophan, the genetic code is degenerated, meaning that all other amino acids can be encoded by at least two and up to six codons [9]. This degeneracy provides nature with a large number of possible permutations to encode each specific protein, giving rise to synonymous variations

through mutations and to primary structure evolution of genes through natural selection of these variants [8].

A first study, in 1981, identified biased codons in 161 full or partial mRNA sequences contained in a nucleic acid database [10]. Following, Ikemura, T. (1985) found that the alternative synonymous codons for each amino acid were not used randomly [14]. Today it is well established that different organisms and genes, tend to use different sets of synonymous codons and the frequency of the phenomenon is biased throughout the protein coding system [11]. Also established is that, while some sites are more likely to be biased, others are well conserved throughout many species, reflecting the action of two evolutionary forces on the genetic code [8], selection and mutation. According to the selectionist theory, codon bias contributes to the efficiency and/or the accuracy of protein expression and is a result of selection. By contrast, the mutational or neutral theory posits that codon bias exists because mutational patterns are not random; i.e. some codons are more mutable and thus have lower equilibrium frequencies, which leads to differences in the patterns of codon biases across organisms [12].

Sharp et al (1987), showed a clear correlation between codon bias and gene expression in *Escherichia coli* and *Saccharomyces cerevisiae* [13]. To quantify the degree of codon bias, several indices were proposed [14-16]. The codon adaptation index (CAI) combined some of these approaches and enabled convenient comparison between different species, which explains its wide use today.

CAI shows, how closely a gene conforms with the codon usage of highly expressed housekeeping genes [13], and thus can be used as a predictor for protein expression. A gene that has a CAI equal to 1 (maximal CAI) means, that it uses only the most frequent codons to encode each of its amino acids. A bias in codon usage can greatly influence: “elongation speed, translation accuracy and fidelity improvement of processes down-stream of translation” [6, 17-18]. However, CAI maximization alone is often not sufficient to achieve high levels of expression of a functional protein and can result in abnormal and none active proteins. At least partially, this can be explained by the role of rare codons, which will be introduced in the following.

A more recently discovered feature of the mRNA primary structure are functionally relevant rare codons, which directly influence the expression dynamics of many proteins. Rare codons are low usage codons, which are paired with low abundance tRNAs. Since the translational elongation rate is tRNA-concentration

dependent, these codons slow down the speed of translation [19-20]. Thus, the translation rate of codons that are read by abundant cognate tRNAs is faster than that of codons read by rare tRNAs. **Rare codon rich regions (RCRRs)** can provide an important time delay at positions that are directly related with co-translational protein folding, and ribosomal traffic control. These positions are often situated in loop and linker regions and several studies could show that only discontinuous translation at these positions enables the proper sequential folding of defined portions of the nascent polypeptide, emerging from the ribosome [19-21]. Recent data suggest that accumulations of rare codons at the 5´end of ORFs can have another important effect on translation dynamics. Those rare codons create a slow "ramp" phase during the beginning of translation that reduces ribosomal traffic jams in a later stage and, as a result, minimize protein expression costs for the cell [22].

Also recently identified was a codon correlation effect in *Saccharomyces cerevisiae*. This means that, after a particular codon is used, the subsequent occurrences of the same amino acid do not use codons randomly, but favour the ones which use the same tRNA [6]. The authors suggest, that codon correlation could either be the result of tRNA diffusion away from the ribosome being slower than translation and/or that some sort of tRNA channelling takes place at the ribosome.

Despite these new insights into the translation process, the relationships between codon usage and gene expression are still not understood in their entireness and one has not arrived yet at general rules of mRNA primary structure and their effects on the translational process.

1.2.2 Codon context

Codon context – the nucleotides surrounding a codon [23] - is another important feature of mRNA primary structure. The combination of some codon neighbours instead of others can affect translation efficiency, accuracy as well as mis- and nonsense suppression [24]. Furthermore, if surrounding nucleotides form mononucleotide repeats, also transcription and translation slippage can occur [11].

During the elongation step, two codons and consequently two tRNAs need to be simultaneously interlinked with the A and P ribosomal sites. Not all combinations of codons and respective tRNAs are equally favourable to interact with the ribosomal

surface because of physical interactions between the tRNA isoacceptors [11, 18]. Others suggest that this could have been the driving force of codon usage evolution [25].

Context biases are present in all three domains of life and are strongest in the nucleotides following the codon in the 3' direction. A recent comparative analysis of 138 organisms, including bacteria, archaea and eukaryotes, even suggests that certain codon context patterns are strongly conserved in a large number of organisms [11].

The codon context in bacteria and archaea appears to be mainly influenced by the translation machinery, while in eukaryotes, context bias seems to be more related to DNA methylation and tri-nucleotide repeats, which are present at higher frequencies [26]. Particularly in fungal species, it was recently shown that codon-triplet context is highly biased and the strongest bias was identified in *Candida albicans* [27], an opportunistic human pathogen. In addition, other studies affirm that the codon-triplets present on the A-, P- and E-sites of the ribosome are determinants for mRNA translation accuracy and efficiency [11, 27]. For instance, the E-site occupation seems to influence the decoding fidelity in the A-site.

Recently, it was suggested that codon context has co-evolved with the structure and abundance of tRNA isoacceptors in order to control translation rates [18] and that codon context might influence translation rates more than single codon usage [11].

1.2.3 mRNA secondary structure

The mRNA secondary structure in the 5'UTR regulates mRNA degradation by specific RNases and thus plays an important role in transcript stability and mRNA half-life. [28-29]. Other mRNA structures can have strong negative effects on translation rate [2, 30-31]. Those effects are caused by their interference with ribosomal binding- and translation initiation sites and depends greatly on the strength and type of mRNA secondary structure [2]. One particularly stable type of structures are pseudoknots, containing at least two stem-loops. The current understanding is that these structures, when very stable, can occlude the ribosomal binding site (RBS), and/or the start codon from the ribosomal machinery.

A recent study of a human protein, expressed in *E. coli* showed that good exposure of the initiation codon to the ribosome improved translation rate; in this particular case 10-fold compared to the wild type [7]. Therefore, a careful and systematic analysis is needed when optimizing mRNA secondary structures. Algorithms

to design RBSs with enhanced affinity to the ribosome are in place and could be shown to improve initiation rate and protein synthesis [32]. However, it has to be kept in mind that secondary structures change dynamically while the ribosome moves along the mRNA and further *in vivo* studies are needed to better understand the effect of dynamic mRNA secondary structure changes on the translation process [2].

1.3 tRNA and codon decoding

Translation fidelity is crucial to guarantee that the mRNA coding sequence is correctly expressed into the respective protein. Therefore, decoding represents a crucial step, because it is the stage where each codon has to be properly associated to an amino acid [33].

Tranfer RNA (tRNA) is the central component of this process [33-34] because it is responsible for establishing the link between the mRNA sequence and the amino acid sequence. The tRNAs are small RNA molecules (70 to 95 nucleotides) that, during the elongation step of translation, have the capacity of reading the coding sequence and transferring the respective amino acids to the growing polypeptide chain.

They have a *cloverleaf* shaped structure (fig. 4), composed of three main regions that are assembled by self-complementation (T-arm, D-arm, Anticodon arm) and an acceptor stem, which attaches to an amino acid by its 3'-terminal site. Each tRNA contains a three base region (anticodon) located in the anticodon arm, which can bind to the corresponding codon on the mRNA chain.

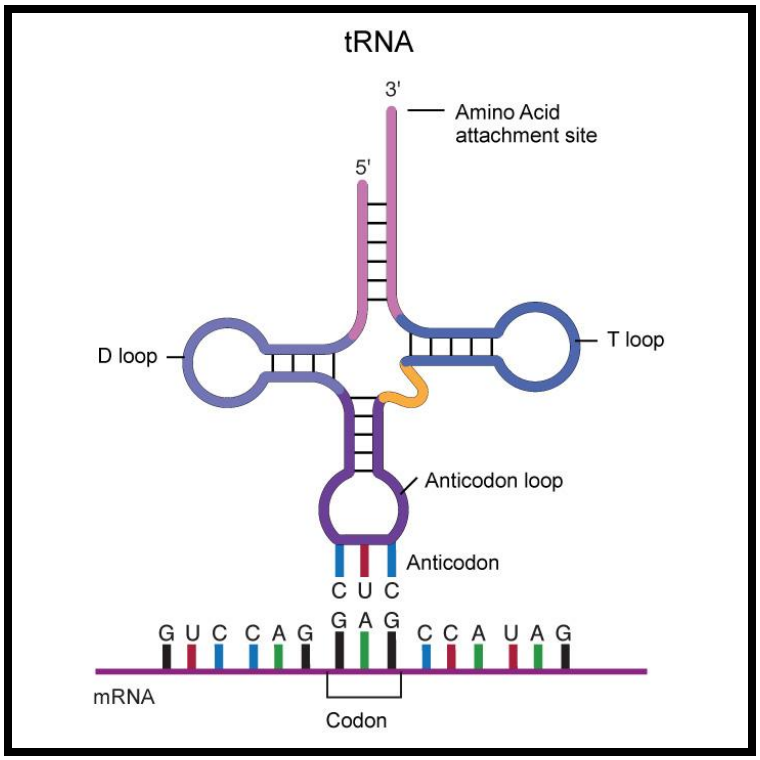


Figure 4. tRNA Cloverleaf Structure. (*NHGRI; www.genome.gov)*

The amino acid incorporation into the tRNA molecule is mediated by a two step reaction called **aminoacylation** [35].

1. Amino acid + ATP → aminoacyl-AMP + PPi
2. Aminoacyl-AMP + tRNA → aminoacyl-tRNA + AMP

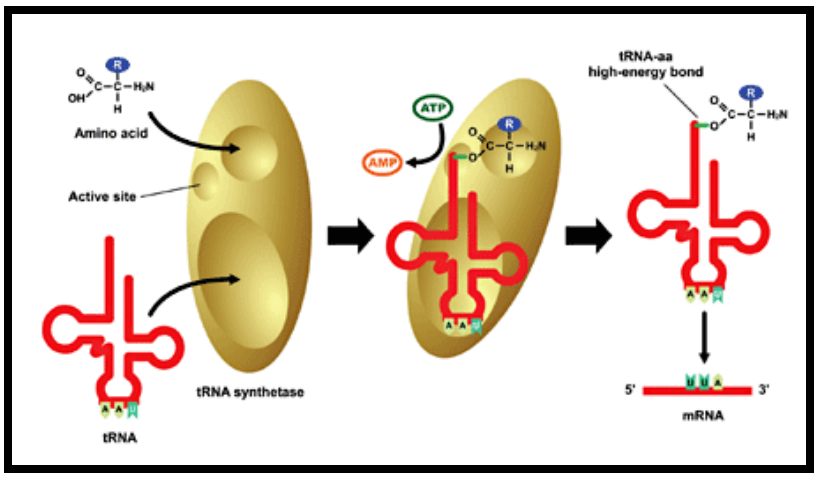


Figure 5. Scheme of the aminoacylation process. [36]

In the first step of the aminoacylation process, an amino acid is activated with ATP. An aminoacyl-tRNA synthetase (aaRS) attaches the carboxyl group of the amino acid to the phosphoryl group of the AMP, and finally produces an aminoacyl adenylate. The activated amino acid is then transferred to the tRNA and the final product, the aminoacyl-tRNA, is released.

The majority of amino acids are esterified onto tRNAs directly by aminoacyl-tRNA synthetases (aaRSs), but in some cases the process can follow indirect pathways, as for Gln, Asn, Cys, and Sec [35].

As described before, there are more sense codons than available amino acids. Since one specific tRNA molecule can only be attached to one single amino acid, and to keep a one-to-one correspondence between codons and tRNAs, 61 different tRNA molecules would be needed. However, some tRNA molecules have the capacity to read multiple synonymous codons, reducing the number of required tRNAs and consequently, tRNA complexity [6]. These multivalent tRNAs use non-Watson-Crick base pairing to recognize synonymous codons and compete between each other until the most stable codon-anticodon connection is established.

Through the advances in genome sequencing, more information is starting to accumulate about tRNA numbers and their organization in different species [34]. The tRNA-isoacceptor number varies between different genomes and generally correlates with variations in codon usage. This connection between tRNA-isoacceptors and codon usage was first identified in *E. coli* and yeast, and since has also been found in other pro- and eukaryotic species [14, 16, 37].

1.4 Translation rate / Protein folding

Frequent codons tend to dominate in highly expressed genes and are normally decoded by abundant tRNAs. Low usage codons are generally decoded by tRNAs of low abundance and consequently their translation rate is slower. In fact, the asymmetric abundance of tRNAs, decoding different sets of synonymous codons, causes a discontinuous translation rate which is distinct among different organisms, tissues and stages of differentiation [20].

Recent data showed that discontinuous elongation rates are closely related to secondary-structural elements in proteins: β-sheets, loops and disordered structures are normally encoded by rare codons, whereas α-helices are encoded by more frequent codons [38]. Further, these data suggest that attenuations at specific sites enable the definition of distinct elements of the nascent protein chain emerging from the ribosome and that translation speed variations can be crucial in the synchronization of the folding process, affecting the final protein conformation. Thus, instructions to coordinate the native three-dimensional protein structure can be encrypted as an additional layer of information contained in the mRNA sequence.

Several algorithms were recently developed to predict and map the folding status of a specific ORF [20, 38-39]. These systems identified putative attenuation regions in prokaryotic and eukaryotic protein sequences, based on factors like: tRNA concentration, codon specificity, tRNA recharging, steric effects and local mRNA secondary structures.

During the folding process, multiple pathways, intermediates and aiding proteins (e.g. the Group I chaperonine complex GroEL/GroES in *E. coli* and peptiyl-prolyl cis-trans isomerases) guide the polypeptide chain to a native state where a free energy minimum is reached [40]. Co-translational folding is the formation of structures as soon as a portion of the nascent polypeptide chain emerges from the ribosomal tunnel during protein synthesis. Particularly, α-helices can be formed and stabilized already inside that tunnel, which occludes approximately 30 nucleotides and is not an absolutely rigid structure [41]. From a thermodynamic perspective, co-translational folding is a favorable process during protein synthesis. As mentioned before, the location of pause sites is normally related with a domain terminus and/or boundaries (loops e.g.), which appears to provide a crucial time delay for the nascent protein to fold correctly.

Considering these findings, it becomes clear that the synthesis of a protein in a heterologous system, with its different codon usage and/or different tRNA pools, must have entirely different translation and folding kinetics than in the native species, with consequences for the final protein conformation and activity.

To summarize, the optimisation of synonymous codons along the mRNA sequence determines whether translation kinetics in the heterologous host occur in the same way as in the native host. For example, a study on *Plasmodium falciparum*

showed that, replacing rare synonymous codons at specific sites, increased the yield and solubility of three recombinant MSP142 proteins expressed in *E. coli* [39].

1.5 Gene optimisation for heterologous protein expression

The optimisation of genes is important for many applications in biotechnology and molecular biology [42], including: gene therapy, DNA vaccination vector production, molecular engineering or heterologous expression [2].

Gene optimisation is the rewriting of an open reading frame (ORF) according to certain gene design rules without changing the encoded protein. This is only possible, and at the same time only necessary, because of genetic code degeneracy. Optimisation of codon usage is one of the most routinely performed optimisation strategies. For this purpose, host specific low usage codons are replaced by host specific frequent codons, which increases the protein production rate. Other factors that influence protein expression, such as mRNA secondary structure, are also considered more and more in gene optimisation. Less common, but also fast advancing, are gene optimisation strategies that improve protein solubility and minimize protein aggregates. These strategies are based on improving the ability of the foreign host to recognize additional instructions regarding translation dynamics that can be contained in the foreign mRNA.

Popular host organisms for heterologous protein expression include: *Pichia pastoris, Escherichia coli, Saccharomyces cerevisiae, Pseudomonas fluorescens* and *Aspergillus phytae*. The selection of a suitable host and the optimisation of production conditions remain important steps in the process and consideration has to be given also to issues like toxicity problems and protease degradation [2].

Gene optimisation tools, such as *Codon Optimiser* [43] and others, allow the manipulation of various gene design parameters. Most commonly, these programs use an algorithm, which finds the most commonly used codons in the host species, and designs an optimised DNA sequence to be expressed (relative synonymous codon usage (RSCU)- or CAI optimisation) [44-45]. Besides codon usage, several other factors can be analysed by these algorithms: codon context, restriction enzyme sites, secondary structure elements, water contact information and GC content [44-45]. However, it is difficult to prioritize these parameters because variables may not be independent of each other.

For this study a codon optimisation software, called ANACONDA®. That software analyses gene primary structures and allows the identification of low usage codons, the determination of CAI, codon context and nucleotide repetitions within entire ORFeomes.

The optimised sequences, bioinformatics tools provide, are then often used for ***de novo* gene synthesis**. Synthetic gene construction allows the production of sequences, without DNA template being required. This can be very useful if the target sequence cannot be easily obtained and if considerable changes need to be made to optimise the sequence [46].

A synthetic gene is normally constructed by ´multiple oligonucleotides assembly´ using methods such as enzymatic ligation, serial cloning or PCR extension [46-47]. The PCR based two-step DNA synthesis (fig. 6) is currently the most widely used. In this method, oligonucleotides are designed to cover both strands of the desired gene, and during a first PCR reaction, the sequence is progressively produced by overlap extension. A second PCR reaction, using two outside primers, is then used to amplify the full-length sequence. However, PCR-based methods are influenced by many factors, and the quality of the final product depends on the number of PCR cycles, gene length, and DNA polymerase fidelity [46].

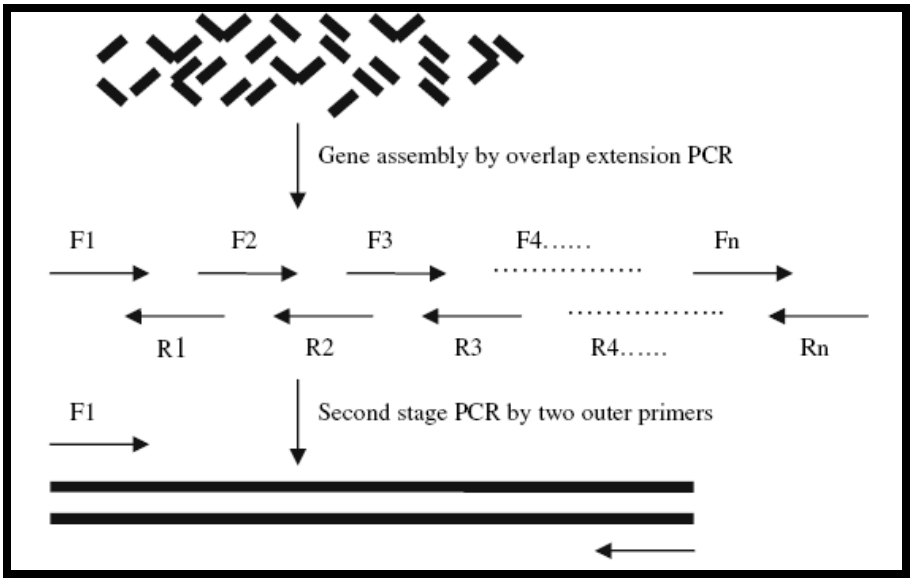


Figure 6. Schematic diagram of PCR-based two-step gene synthesis. Adapted from [48].

To correct errors in final sequences and for small modifications, often **site-directed mutagenesis (SDM**) is used [49]. This molecular technique is based on PCR amplification and is performed to create site-specific mutations. The procedure requires a primer, complementary to the DNA template, but containing an internal mismatch to create the desired mutation. This oligonucleotide is then hybridized within the target gene and a polymerase enzyme performs the extension of the sequence.

In summary, with the improvements in gene design and gene synthesis techniques, it is now possible to synthesize optimised gene sequences of up to 25 kb by fully automated procedures [47]. Furthermore, several prokaryotic and eukaryotic heterologous protein expression systems are established, providing a choice of selecting the best-suited host organism for the specific needs. However, gene optimisation rules for heterologous protein expression are still not well enough understood to routinely optimise any gene for any host. In fact, the many layers of interdependent factors and the uniqueness of translation dynamics, required to produce a certain protein, may prevent researchers from developing reliable algorithms based on standard criteria for some time to come. Also, it seems that bioinformatics approaches are getting ahead of laboratory testing and many of the patterns that were discovered *in silico* still need to be validated in robust biological models.

2 Project Outline

As part of Mephitis, a EU funded project to develop novel drugs against the protein synthesis machinery of *P. falciparum*, the RNA biology lab is involved in the study of novel approaches to improve the heterologous expression of plasmodial proteins. Improving the heterologous expression of proteins from this malaria-causing parasite is a pressing issue because ORFeome structure incompatibilities often impair protein synthesis, resulting in low expression levels and/or misfolded, insoluble and nonfunctional proteins. As a result, structural information on plasmodial proteins is still scarce, which ultimately also limits drug design against specific targets.

Previous work in the RNA biology lab focused on studying the effects of CAI, context and CAI/context optimisations on expression and protein solubility [50]. As a model, a *P. falciparum* lysyl-tRNA synthetase (*Pf* LysRS) was expressed in *E. coli*. The same model was employed to test the effects of single RCRRs on protein expression and solubility. The focus of the present work was to further study the functional relevance of rare codons in the *Pf* LysRS, to include codon harmonisation in the optimisation process and to widen the optimisation efforts to other proteins in the search for gene primary structure rules.

More specifically, the objectives of this project were:

1) to study the effects of increasing numbers of RCRRS in a CAI-optimised LysRS on protein expression and solubility. For this, synonymous genes with increasing numbers of RCRRS were created by sequentially introducing rare codons at specific positions in the LysRS gene using PCR-based site directed mutagenesis. Overexpression analysis of both soluble and insoluble protein fractions was performed to study if the introduced rare codons influenced protein solubility; an expected effect of changed protein folding dynamics.

2) to examine whether protein solubility of the LysRS can be improved by codon harmonisation. For this, a novel gene optimisation tool was employed to harmonise the *Pf* LysRS gene for its expression in *E. coli* and the harmonised gene was obtained from a commercial provider. To study the expression dynamics of the codon-harmonised gene, overexpression analysis was again performed on both soluble and insoluble protein fractions.

3) to test the ability of the different synonymous *Pf* LysRS genes in rescuing a temperature sensitive *E. coli* strain, deficient in its natural LysRS. For this, a temperature sensitive strain was transfected with the different gene constructs and survival of the strain was assessed at elevated temperatures. This functional assay was intended to give additional information on the protein folding status and the functionality of the plasmodial LysRS.

4) to study the effects of codon context on translational accuracy by de-optimizing the codon context of a native *E. coli* gene. For this approach, a reporter system based on the *E. coli* β–galactosidase protein lacZα was chosen. Because in this system the functional β–galactosidase is composed by two separate peptides, LacZα (encoded in a plasmid) and LacZΩ (encoded in the genomic DNA), manipulating the α-peptide is comparatively simple and allows the study of amino acid misincorporation through a functional heat stability assay.

3 Materials and methods

3.1 Bacterial strains and plasmids

3.1.1 Bacterial strains

Standard cloning and screening procedures were performed using the ***E. coli* strain JM109** (*end*A1, *rec*A1, *gyr*A96, *thi*, *hsd*R17 (r_k^-, m_k^+), *rel*A1, *sup*E44, Δ(*lac-pro*AB), [F´ *tra*D36, *pro*AB, *laq*I^qZΔM15]) (Yanisch-Perron et al 1985) and ***E. coli* strain DH5-α** (*fhu*A2 Δ(*arg*F-*lac*Z)U169 *pho*A *gln*V44 Φ80 Δ(*lac*Z)M15 *gyr*A96 recA1 *rel*A1 *end*A1 *thi*-1 *hsd*R17).

E. coli strain JM109: The *endA1* mutation of this strain leads to an improved yield and quality of isolated plasmid DNA and the *recA1* genotype prevents recombination with host chromosomal DNA and improves plasmid stability. Other strain features include reduced cleavage of heterologous DNA by endonuclease (*hsd*R17) and the possibility for blue/white screening through an F' episome carrying Δ(*lacZ*)M15.

E. coli strain DH5-α*:* The endA1 mutation of this strain inactivates an intracellular endonuclease which degrades plasmid DNA in many miniprep methods and *recA* eliminates homologous recombination. Other features include blue-white screening with *lac*Z based vectors - Δ(*lac*Z)M15 - and an amber suppressor (*gln*Y44).

For protein overexpression analyses, the ***E. coli* BL21 (DE3) strain** was used. This strain is suitable for high-level protein expression using T7 promoter-driven vectors such as the pET vectors. The genotype of this strain is: F-, *omp*T, *hsd*S(r_B^-, m_B^-), *gal*, *dcm*, λDE3 (*lac*I, *lac*UV5-T7 gene 1, *ind*1, *sam*7, *nin*5). Strain features include the lack of two key proteases (*lon* and *ompT*), that otherwise could degrade recombinant protein and restriction deficiency ($hsdS_B^-$), protecting introduced plasmids from degradation. **KRX *E. coli*** cells were used to provide efficient transformation and tightly controlled protein expression at the same time. Single Step (KRX) is an *E. coli K* strain that contains a chromosomal copy of the T7 RNA polymerase driven by a rhamnose promoter (*rha*BAD) to provide effective control of the proteins expressed via a T7 promoter. Genotype: [F-prime, *tra*D36, *-delta-omp*P, *pro*A+B+, *lac*I^q, *-delta-*

(*lac*Z)M15] -*delta*-*omp*T, *end*A1, *rec*A1, *gyr*A96 (Nalr), *thi*-1, *hsd*R17 (r_k-, m_k+), e14-(McrA-), *rel*A1, *sup*E44, -*delta*-(*lac*-*pro*AB), -*delta*-(*rha*BAD)::T7 RNA polymerase.

The ***E. coli* PALΔSΔUTR strain** (F−(*lac*-*pro*) *gyrA rpoB metB argE*(Am) *ara suPf* Δ*lysS*::*kan* Δ*lysU srl*-300::Tn10 *recA*56 (**pMAK705 *lysU*$^+$**)) is a temperature sensitive strain, obtained from the group of Dr. Lluis Ribas de Pouplana at the *Fundació Parc Científic de Barcelona*, in Barcelona. The strain was derived from PAL3103SK, introducing a deletion in the chromosomal *lys*U gene by recombination with a homologous sequence introduced into the temperature-sensitive plasmid pMAK705. Concomitantly, the intact *lys*U gene was recovered in plasmid pMAK705. The strain was then made deficient in recombination by transducing the recA56 allele of the Hfr strain JC10240. In the present work, this strain was used to test the ability of *Pf* LysRS activity on rescuing PALΔSΔUTR at 42 ºC.

3.1.2 Plasmids

The **plasmid pET19b** (see Appendix A) is a cloning and expression vector of 5,7 kb, which has a origin of replication for *E. coli*, a lacI coding sequence and a multiple cloning site (MCS) where genes of interest can be inserted. This low-copy vector carries an N-terminal His•Tag® sequence (10x His) followed by an enterokinase site. In preparation for this study, a derivative of the pET19b vector was created that had the original His-tag and the enterokinase site replaced by a shorter His-tag (6x His) and a Flag-tag.

The **plasmid pUC19** is a cloning vector (see Appendix A) commonly used in *E. coli*. It is a small plasmid (2686 base pairs) with high copy number and it carries a MCS that contains unique sites for 13 restriction endonucleases. The MCS is in frame with the lacZα gene, which can be induced by IPTG and enables the screening for insertions using α-complementation. Through α-complementation, functionality of a defective form of the β-galactosidase enzyme, encoded by host genome, is restored (mutation Δ*(lacZ)M15*). In this work, pUC19 was used for the context de-optimisation of the β-gal α-peptide.

3.2 Growth medium

The growth medium used was Luria-Bertani (LB), a rich medium used for growing bacteria like *E. coli*. It is composed of 1.0 % Tryptone, 0.5 % Yeast Extract and 1 % Sodium Chloride (protocols for the preparation of LB medium and LB agar plates are shown in Appendix E).

3.3 Cloning and transformation

3.3.1 Insertion of genes into cloning and expression vectors

The plasmids containing the target genes and vectors (pET19b or pUC19) were first digested with restriction enzymes. In particular, pET19b and LysRS constructs were double digested with NcoI and XhoI, for 2 hours at 37 ºC. pUC19 vector was first digested with XhoI for 2 hours at 37ºC and after a purification step, either with a PCR clean up or a gel purification kit (Qiagen), was digested with NedI for 2 hours at 37ºC (for details see Appendix E).

Following, the restriction enzymes were heat inactivated for 20 min at 80 ºC. The digested vectors were then treated with shrimp alkaline phosphatase (SAP) to dephosphorylate the 5'-ends and both the target genes and the cut vector were purified either with a PCR clean up or a gel purification kit (Qiagen).

Ligations of target genes into the vector were performed overnight at 16 ºC, using T4 Ligase (New England Biolabs), followed by heat inactivation of the enzyme at 65 ºC, for 10 min.

3.3.2 *E. coli* transformation

Depending on the purpose of transformations, different *E. coli* strains were used. For general cloning procedures the strains JM109 and DH5α were used, whereas the strains BL21 (D3) and KRX (Promega) were used for the overexpression of proteins. With the exception of the KRX cells, which were obtained as competent cells, the other strains were made competent in the lab using a protocol for chemically preparing competent cells (Appendix E for detailed protocol). Before transforming competent

cells, they were defrosted on ice. Once thawed, transformations were initiated by adding 50 ng of plasmidic DNA to 200 µl of competent cells. The DNA and cells were mixed gently and incubated on ice for 30 min. Following, the cells were heat shocked for 90 seconds at 42 ºC and then cooled down on ice for 2 minutes. Then, 800 µl of cold SOC medium (recipe E) were added and cells were incubated for one hour at 37 ºC and shaking at 180 rpm, allowing them to recover. The cells were then gently pelleted for 2 min at 2500 rpm in a table–top centrifuge and approx. 50 µl of the supernatant was removed. The remaining medium was used to resuspend the cells before they were plated out on LB agar plates, containing ampicillin at a concentration of 75µg/ml. The plates were incubated overnight at 37 ºC and then stored at 4 ºC until further use.

3.4 PCR-based methods

3.4.1 Background on the polymerase chain reaction (PCR)

The polymerase chain reaction (PCR) consists of a repetitive series (typically 30-40 cycles) of three fundamentals steps: In the first step, the double-stranded DNA template is denatured by increasing the temperature to ~95 ºC. At this temperature, the hydrogen bonds between the complementary bases break up, yielding single DNA strands. Following this step, the annealing occurs at a temperature below the melting temperature (Tm) of two oligonucleotide primers (usually 50–65 °C), allowing them to bind to the complementary regions of the single-stranded DNA template. During the third step, extension or elongation, the *Taq* DNA polymerase recognizes the short double stranded sections of DNA created by the annealed primers and extends the primers in 5´to 3´direction by incorporating dNTPs. The optimal temperature for this enzymatic reaction is around 72 ºC. The newly created double stranded DNA molecules then serve as templates for the next cycle of denaturing, annealing and elongation.

3.4.2 Colony PCR amplification

The use of intact cells from a colony as template for a PCR (colony PCR) is a method that allows the rapid screening of clones to identify the presence of a target DNA molecule. The crucial step in this kind of PCR is to make the DNA of the cells available for amplification. This is usually achieved by boiling the cells in MQ water, which breaks the cells and sets the DNA free. The PCR components and cycling protocols are generally the same as for purified DNA templates.

In the present study, individual colonies were picked using sterile P10 micropipette tips. The picked cells were transferred onto a grid on LB agar plates to establish a clone library for later plasmid extractions and the remaining cells on the tips were transferred into 0,2 ml PCR reaction tubes, containing 5 µl of MQ water. The tubes were then incubated for 5 min at 95 ºC and centrifuged for 1 min at 16.100 g in a microcentrifuge (5418 R, Eppendorf) to pellet cell debris. The template (1 µl of the supernatant) was transferred into a new PCR reaction tube and the remaining PCR components were added in form of a master mix. For the master mix, multiples of 0,125 µl *Taq* DNA polymerase (1 U/µl), 2,5 µl of 10 X Buffer (500 mM KCl, 100 mM Tris-HCl (pH 9.0), 1.0% Triton X 100), 0,3 µl of dNTPs (5 mM) and 0,125 µl of each primer (10 µM) were combined and brought to a final volume of 24 µl per PCR reaction with sterile MQ water. The PCR protocol consisted in 35 cycles, of denaturing at 94 ºC for 60´, annealing at 55 ºC for 90´ and extension at 72 ºC for 60´ in a MyCycler™ thermal cycler (BIORAD). The PCR products were separated for 30-45 min at a constant voltage of 80 V on 1,2 % agarose gels alongside a DNA ladder (GeneRuler^{TM} 100bp DNA ladder Plus, Fermentas) and colonies that resulted in the expected amplicons were identified.

3.4.3 SDM – Site directed mutagenesis

In vitro site-directed mutagenesis was used to generate modified DNA sequences containing mutated codons. The technique uses supercoiled double stranded DNA (dsDNA) as template and two synthetic oligonucleotide primers carrying a specific mutation, each one complementary to opposite strands of the vector. After annealing, the primers are extended by the *Pf* uTurbo polymerase and incorporated in the newly formed DNA sequence. Following the PCR (typically, 16-18 cycles), the nicked

mutated plasmid is separated from the parental DNA template, through digestion with the endonuclease, *Dpn* I. This restriction enzyme specifically cleaves methylated DNA and therefore, it digests the template plasmid, without affecting the PCR product.

Using SDM, specific mutations were sequentially introduced in the CAI-optimised LysRS sequence, corresponding to nine rare codon rich regions, previously targeted (see Appendix B).

The protocol was carried out based on the QuickChange Site-Directed Mutagenesis Kit (Stratagene). The template DNA was prepared at 2,5; 10 and 20 ng/µl and each reaction of 25 µl contained 5 mM dNTP mix, 10 µM of each primer, 10× Reaction buffer, 2,5 U/µl *Taq* DNA polymerase. PCR amplifications were performed using a protocol of 18 cycles, with denaturing at 95 ºC for 30´´, annealing at 55 ºC for 1´ and extension at 68 ºC for 8´ in a MyCycler^{TM} thermal cycler (BIORAD). After that, the PCR reactions were incubated with 10 U of *Dpn* I.

3.4.4 Agarose gel electrophoresis

At neutral pH, DNA is negatively charged and in an electric field migrates from the negative to the positive pole. When this migration is forced through a suitable matrix, shorter molecules move faster than longer ones because they can migrate more easily through the matrix. Agarose gel electrophoresis is a method used to separate DNA and RNA fragments by length and to assess their quality and yield. Agarose gels are generally made between 0.7% and 2% of agarose. A gel with a lower agarose percentage permits a better resolution of large DNA fragments (5-10kb), while a gel with a higher agarose percentage is more appropriated to separate small fragments.

After separation, the DNA can be visualised in the gel by e.g. ethidium bromide. Ethidium bromide is a fluorescent compound that binds strongly to double stranded DNA by intercalating between the bases. When intercalated, it strongly absorbs UV light and emits visible orange light. The length of the DNA fragment of interest is then determined by comparison with a DNA marker of known fragment lengths.

3.4.5 PCR purification

To purify PCR amplicons, a QIAquick PCR Purification Kit from Qiagen was used. The kit works on the basis of reversible adsorption of nucleic acids to silica-gel, in the presence of high concentrations of chaotropic salts and high pH. Chaotropic salts (e.g. urea, thiourea) disrupt hydrogen bond structures in water, affect the nucleic acids secondary structure and decrease the solubility of DNA in water.

The first step of the protocol consisted in adding the binding buffer (PB; pH 7.5 and containing the chaotropic salt guanidine hydrochloride) to the samples. Then, the DNA solution was passed through the silica-gel membrane by centrifugation to bind the DNA, followed by a washing step with an ethanol-containing buffer (PE buffer) to remove salts. Finally, the DNA was eluted in a low-salt solution (EB buffer; 10 mM Tris·Cl, pH 8.5) (Detailed protocol in Appendix E).

3.4.6 Spectrophotometric quantification and quality analysis of DNA

Following the preparation of DNA and prior to most downstream applications, it is important to assess the quantity and quality of DNA. The two most commonly used methods to quantify DNA are: gel electrophoresis and spectrophotometric analysis. The spectrophotometric analysis is based on the UV absorption properties of nucleic acids and potential contaminants such as proteins.

Nucleic acids have an absorption maximum around 260nm, whereas proteins have a maximum absorption around 280nm. Based on the absorption at 260nm, the amount of DNA can be quantified using the formula:

DNA concentration (ng/µl) = $\underline{OD_{260} \text{ x (dilution factor) x conversion factor}}$

The spectrophotometric conversion factors for nucleic acids are:

double stranded DNA: 1 OD_{260} = 50 ng/µl

single stranded DNA: 1 OD_{260} = 33 ng/µl

single stranded RNA: 1 OD_{260} = 40 ng/µl

The ratio of OD_{260}/OD_{280} gives an indication for the purity of the sample. A ratio between 1,8-2,0 indicates that the nucleic acid in the sample is pure; a ratio lower than 1,8 indicates protein impurities and a ratio higher than 2,0 indicates a chloroform or

phenol contamination. Compared to the gel electrophoretic analysis, one disadvantage of the spectrophotometric analysis is that sheared or otherwise degraded DNA is indistinguishable from high quality DNA, since both have equal UV absorption.

In the present work, DNA quantifications and quality analyses were performed spectrophotometrically, using a NanoDrop™ 1000 spectrophotometer (Thermo Scientific). This type of spectrophotometer allows the analysis of very small sample volumes and does not require cuvettes (Fig. 7). For protocol details see Appendix E.

Figure 7. NanoDrop™ 1000 spectrophotometer (Thermo Scientific)

3.4.7 Preparation of samples for sequencing

Once positive clones were putatively identified by colony PCR and gel electrophoretic analysis, their plasmids were extracted using a miniprep kit (GeneJET™, Fermentas) and the plasmids were sequenced to confirm identities. For this, 50 ml of liquid LB medium were inoculated with the respective clones and grown overnight at 37 ºC and shaking at 180 rpm. The following day, the cultures were first cooled on ice and then pelleted for 10 min at 3220 *g* and 4 ºC in a refrigerated centrifuge (5810R, Eppendorf). The pelleted cells were resuspended in a SDS/alkaline lysis buffer (composed of salts, detergents and RNase A) to break the cells and degrade contaminating RNA. Next, the solution was neutralized to allow the DNA binding to the silica membrane in spin columns. After several washes to remove contaminants, the DNA was eluted in a small volume (≤50 µl) of elution buffer. Before the plasmid DNA was stored at -20 ºC, the DNA concentration and purity was measured using a

Nanodrop™ 1000 (Thermo Scientific). The sequencing of extracted plasmids was performed by a commercial sequencing provider (STAB VIDA) using either standard or custom primers.

3.5 Heterologous expression of protein and overexpression analysis

For the heterologous expression of the different gene constructs, they were transfected into either BL21 (DE3) or KRX cells (see section 3.3.2 for details). The used T7 promoter expression system is capable of producing higher protein yields than any other bacterial expression system. Depending on the E. coli strain, protein expression was induced by IPTG or rhamnose. Subsequent to induction, three time points were sampled, cells were disrupted by sonication (sonicator from IKA LaboraTechnik), soluble and insoluble protein fractions were extracted and protein concentrations were quantified using a BCA™ Protein Assay Kit from Pierce.

3.5.1 Induction

In preparation for the overexpression of proteins, 15 ml pre-cultures (LB medium with ampicillin at a concentration of 75 µg/ml) were inoculated with single colonies of transformed *E. coli* strains and grown overnight at 37 ºC with shaking at 180 rpm. The following day, 4 ml of pre-culture were transferred to a 250 ml Erlenmeyer flask containing 100 ml of new LB medium and the same concentration of ampicilin. The remaining pre-culture was used to extract the plasmid for sequence confirmation as described above. After 2 hours of incubation at 37 ºC with shaking, the OD was measured using a microplate spectrophotometer (iMark, Biorad) at 595 nm. Once the cultures reached an OD of 0,3-0,5, 20 ml of culture were transferred to a 50 ml Falcon tube, pelleted at 3220 *g* in a refrigerated centrifuge (5810R, Eppendorf) at 4 ºC, the medium was decanted and the pellet was immediately frozen at -20 ºC to avoid protein degradation. In the remaining culture, protein expression was induced by adding either IPTG at a final concentration of 1 mM in case of the BL21 (D3) strain or rhamnose at a final concentration of 0,1% in case of KRX cells. The 20 ml samples were then collected after 1,5 h, 3 h and 4,5 h of induction. Their OD was measured and the cells were pelleted and preserved as described above.

3.5.2 Protein Extraction

For the extraction of soluble proteins, pelleted cells were thawed on ice and resuspended in 1 ml of cold 1x PBS (recipe Appendix F). Keeping the samples on ice, the cells were disrupted by sonicating three times for 10 seconds using a sonication probe. The samples were then transferred to 1.5 ml reaction tubes, centrifuged for 15 min at 16.100 g and 4 ºC and the supernatant, containing the soluble proteins, was collected and stored at -20 ºC. Then 1 ml of resuspension buffer (recipe Appendix F) was added to the pellets and the samples were incubated overnight at 4 ºC. Once the pellets were completely dissolved, the samples were centrifuged again for 15 min at 16.100 g and 4 ºC and the supernatant containing the insoluble proteins was collected and stored at -20 ºC.

3.5.3 Protein quantification

Total protein concentrations in the soluble and insoluble fractions were determined using the BCA™ Protein Assay Kit (Pierce). Protein standards were prepared by a dilution series of bovine serum albumin (BSA) with the following concentrations: 2000 µg/ml, 1500 µg/ml, 1000 µg/ml, 500 µg/ml, 250 µg/ml, 125 µg/ml, 25 µg/ml and 0 µg/ml. The working reagent was prepared by mixing 50 parts of BCA™ reagent A with 1 part of BCA™ reagent B. The reactions were prepared in 96 well plates, each well containing 25 µl of sample or standard and 200 µl of working solution. If protein concentrations turned out to be higher than the highest standard, the assay was repeated with diluted samples. Soluble protein samples were diluted with 1x PBS, insoluble protein samples with solubilisation buffer. After adding the working solution, the samples were mixed for 30 seconds on a vortexer with a 96 well plate adapter and then incubated for 20-30 min at 37 ºC. After the incubation the OD was measured in a microplate spectrophotometer (iMark, Biorad) at 595 nm and protein concentrations were determined based on a linear regression analysis of the protein standard series. All measurements of samples and standards were performed in duplicates.

3.5.4 SDS- PAGE

To separate proteins according to their size, protein extracts were submitted to sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE). SDS, an anionic detergent, is used in SDS-PAGE to reduce proteins to their primary (linearized) structure, destroying non covalent bonds so their negative charge is proportional to their molecular weight. Thus, when an electrical field is applied, proteins will migrate from the negative to the positive pole according to their molecular weight and migration is not influenced by secondary, tertiary or quaternary structure.

Polyacrylamide Gel Preparation

Polyacrylamide is a mixture of two polymers, acrylamide and bisacrilamide. The first one is a linear molecule whereas the second as a "T" shape, and when mixed they form a matrix with different separation gradient, depending on the concentration of each polymer. Each gel is composed by a resolving gel with a percentage of acrylamide adapted to the size of the target protein. The smaller the molecular weight, the higher is the percentage to be used. For the *Pf* LysRS, with a molecular weight of ~69 kDa, 12% resolving gels were used. To study the α-peptide of the β-galactosidase, which has a molecular weight of ~13 kDa, 20% resolving gels were used.

To improve the resolution, proteins first need to pass through a stacking gel on the top of the resolving gel. Its lower pH and acrylamide concentration (4%) as well as the different ionic strength, allows the proteins to be concentrated during the first minutes (~10-15min) of electrophoresis, before entering the resolving portion of the gel.

Table 1. Components and their function in SDS-PAGE

Component	Function
SDS Solution (10%w/v)	Sodium dodecyl sulphate , detergent that denatures and binds to proteins making them evenly negatively charged.
Tris-HCL (1M, pH8)	Buffer that absorbs counter ions (H^+ and OH^-) keeping the solution that in at a stable pH level.
Bis-/acrylamide (40%)	Polymer forming matrix
APS (10mg/100ml) TEMED	Polymerizing agents

The resolving and the stacking gel were prepared with the components described on top but without TEMED. A volume of 2 ml was removed, mixed with 30 µl of TEMED and added to the assembly "cassette" formed by two glasses. The resolving gel with 15 µl of TEMED was added and overlayed with MQ water avoiding dehydration that could interfere with polymerization. After polymerization, the stacking gel was added along with the combs. The assembly was incorporated and filled with 1x SDS Running Buffer (recipe Appendix F).

Sample preparation for SDS-PAGE

Protein samples for SDS-PAGE were prepared by adding 6x loading buffer in a ratio of 1:6, followed by denaturing the samples and a prestained protein marker for one minute at 95 ºC. Then, known amounts of total protein were loaded into the gel pockets and the gels were run at 80-90 V until the samples reached the stacking gel. Then, the voltage was increased to 110-120 V and the gels were run until the dye front reached the bottom of the gel.

3.5.5 Western blotting and immunodetection

In this process, the proteins on the polyacrylamide gel are transferred to a membrane, normally made of nitrocellulose or PVDF (Polyvinylidene Fluoride).

For the transfer a semi–dry blotting system (Trans-Blot®, BIORAD) was used. A stack with the following order from cathode to anode was assembled: three sheets of 3M filter paper soaked in transfer buffer, TGM (recipe Appendix F), gel, nitrocellulose membrane, three sheets of 3M filter paper soaked in transfer buffer. The transfer buffer TGM provides electrical continuity between the electrodes and provides a chemical environment that maintains the solubility of the proteins without preventing the adsorption of the proteins to the membrane during transfer. It is necessary that the membrane is located between the gel and the cathode, as the proteins move towards the positive pole. Once the stack was prepared, it was placed in the transfer system, and a current of 2 A x cm^{-2} was applied for 40-45 minutes. Following the transfer, the membranes were washed twice in 1X TBS for 15 minutes with gentle agitation, to remove any excess of reagents and non-bound protein.

To prevent non-specific interactions between the membrane and the antibody used for detecting the target protein the membranes were saturated ("blocked") with a solution of 3 % non-fat dry milk in Tris-buffered saline (TBS), for 2 hours with gentle agitation. Afterwards, the membranes were washed twice for 5 minutes in 1x TBS (in Appendix F) with gentle agitation.

Detection

Expression of the fusion proteins, containing a Flag- and/or His-tag, was performed using anti-Flag or anti-His antibodies, both raised in mouse.

Following the blocking and washing, the respective primary antibody was added to the membranes in a solution of 3 % non-fat dry milk in 1x TBS. For this, the membranes were placed in heat sealable plastic bags, and each membrane was incubated overnight, at 4 ºC and agitation. After the overnight incubation, the membranes were washed twice for 5 minutes with 1x TBS to remove unbound primary antibody. Then, the membranes were incubated with the secondary antibody, a goat anti-mouse antibody reactive against both primary antibodies and coupled to a fluorochrome that allows subsequent visualization. A 1:10.000 dilution of goat anti-mouse antibody was prepared in 1x TBS and 3% non-fat dry milk. Again, incubations were performed in heat-sealable plastic bags, which were then wrapped in aluminium foil (to avoid fluorochrome degradation by light), at 4 ºC for 1 hour, with agitation. Following, three washes were performed with 1x TBS-Tween (recipe Appendix F), a detergent that washes off unbound antibody from the blot and removes any proteins that are non-specifically bound. The membranes were analysed at 700 and 800 nm using a Odyssey Li-COR fluorescence imager (Bioscience).

3.6 Rescuing assay to test *Pf* LysRS activity in *E. coli*.

The strain PALΔSΔUTR has both chromosomal *lysS* and *lysU* genes (encoding the two *E. coli* LysRSs) disrupted and also contains a temperature-sensitive plasmid (pMAK705) carrying the *lysU* gene. By containing that vector alone (PALΔSΔUTR/pMAK705 *lysU* $^{+}$) its growth is also temperature-sensitive and does not occur at temperatures ~42 °C.

Five pET19b plasmids, each containing a differently optimised, synonymous *Pf* LysRS gene, were used to transform *E. coli* PALΔSΔUTR. The pET19b vectors tested in this functional assay contained the following *Pf* LysRS genes (see Appendix B for detailed sequences):

- Native *Pf* LysRS gene;
- CAI-optimised *Pf* LysRS gene;
- CAI CON-optimised *Pf* LysRS gene;
- Context-optimised *Pf* LysRS gene;
- Harmonised *Pf* LysRS gene.

In the first step, competent PALΔSΔUTR cells were prepared, as well as the LB agar plates containing chloramphenicol (100 µg/ml) and ampicillin (75 µg/ml) to select the transformants that contained both pre-existent (pMAKlysU) and introduced pET19b plasmid, respectively. The *E. coli* PALΔSΔUTR transformation was then performed (see section 4.2. for transformation procedure) for the 5 plasmids described above and a negative control was used (cells without pET19b plasmid). Further, the plates were grown overnight at 30 ºC.

In the following day, tubes with 20 ml of LB medium, containing 75 µg/ml of amplicillin and 100 µg/ml of chloramphenicol, were prepared to test 4 different conditions, for each of the 6 samples:

- 30 ºC without IPTG induction;
- 30 ºC with IPTG induction;
- 42 ºC without IPTG induction;
- 42 ºC with IPTG induction.

Therefore, the LB liquid medium of the 24 tubes (6 strains x 4 conditions) were inoculated with a single colony of each plate already prepared and with or without IPTG (final concentration: 1 mM), for each temperature being tested. The cultures were then grown at 30 ºC or 42 ºC with agitation and the OD_{595} was measured after 0, 1, 4 and 7 hours.

3.7 β-Gal assay

The β-galactosidase was used as a reporter protein to quantify mistranslation induced by context de-optimisation of the *lacZα* gene. Total β-galactosidase activity present in *E. coli* BL21 (DE3) was determined as previously described [51], with some adaptations. In brief, cultures were grown to an A_{595} of 1,0; pelleted by centrifugation at 3220 *g* for 5 min at 4 ºC (Centrifuge 5810R, Eppendorf) and concentrated twice in buffer Z [60 mM Na_2HPO_4, 40 mM NaH_2PO_4, 10 mM KCl, 1 mM $MgSO_4$, pH 7.0]. The cells were then split in seven 2 ml reaction tubes and permeabilized by adding 50 µl of chloroform and 20 µl of 0.1% (w/v) SDS, followed by vortexing for 10 sec. The permeabilized cells were incubated 5 min at 28 ºC (Termomixer Confort, Eppendorf). All tubes, except the controls, were transferred to a 53 ºC water-bath for 2, 4, 8, 12, 20 and 30 min for thermal inactivation. At the respective times, the tubes were collected and put on ice. The fraction of β-galactosidase that remained functional was determined by adding nitrophenyl-β-D-galactosidase (ONPG) at a final concentration of 748 µg/ml and incubating the samples for 16 min at 28 ºC with agitation at 900 rpm (Thermomixer comfort, Eppendorf). The reactions were stopped with 400 mM $NaCO_3$ (710 µl of the 1M $NaCO_3$ stock solution) and the synthesis of *o*-nitrophenol was monitored. For this, the stopped reactions were centrifuged at room temperature for 10 min at 16.100 *g*, 200 µl of the supernatant were transferred to a 96 well plate and the absorbance was measured at 415 nm using a microplate reader (iMark, Biorad).

3.8 Samples preparation for mass spectrometry analysis

3.8.1 In-gel digestion

The purified proteins (in this study, *Pf* LysRS: WT, CAI, CAICON, DIG2, Harmonised and also 3 different truncated proteins, derived from the WT. See fig. 10, section 4.4) were loaded in 7 poliacrylamide gels and those were stained with Comassie blue R-250 (Appendix F for recipe) during 2 hours (Fig. 8) in agitation. Further, the distaining reagent (10% ethanol, 7,5% acetic acid in MQ) was used for 3 hours to take out excess dye and the gels were transferred to a fixation solution (40% methanol, 10% acetic acid in MQ) overnight.

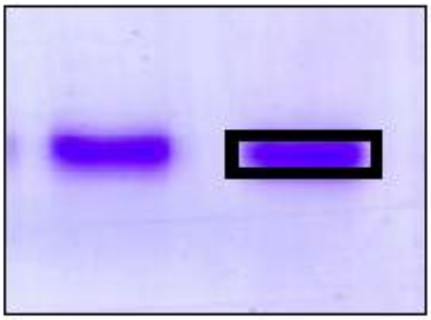


Fig. 8 - Coomassie Blue staining

The bands containing the target proteins were carefully cut in 2 to 3 pieces using a spatula (as shown in fig. 2) and each of them was transferred to a sterile Eppendorf tube. Particular attention was given to not contaminate the samples with keratin from the skin or hair by always using clean gloves and by avoiding contact with the gels.

To start the In-gel digestion, 50 µl of Buf1 (Salt ammonium bicarbonate (NH_4HCO_3) buffer) was added to all the tubes and a 15 min incubation was done at room temperature. Following this step, the supernatant was removed from all the tubes and 50 µl of the solution ACN (Ultra pure acetonitrile) was added to all the samples. After this incubation time, the supernatant was removed again and the procedure was repeated one more time since the Buf1 addition. Further, other 50 µl of ACN were added to the tubes for 10 min to completely desiderate the spots (which already shown a

white colour). That stage was completed by removing the ACN and drying the samples in a Speed Vac for 45 min.

On each tube, a digestion with 25 µl of Trypsin was performed for 60 min, at 37 ºC (to obtain the peptides of each protein) and following this step, 25 µl Buf1 were added on top of the solution for an overnight incubation at the same temperature.

3.8.2 Extraction of the obtained peptides

On the following day, the supernatant was transferred into new tubes and 25 µl of AF (Formic acid (10% v/v)) were added to the remained gel fragments of the previous Eppendorf tubes for 30 min incubation at room temperature. The supernatant was collected again to the new respective tubes and this step was repeated two times using a solution of 25 µl AF + 25 µl ACN. The digests that were collected in the new Eppendorf tubes were combined by protein type (in this project, 7 replicates were collected for each 8 types of proteins loaded in the gels) and put in the Speed Vac to dry for 1 hour.

4 Results

4.1 Effects of RCRRs and codon harmonisation on protein expression and protein solubility of the plasmodial LysRS

Effect of RCRRs:

With the aim to increase the solubility of a CAI-optimised *Pf* LysRS, nine previously identified gene regions were targeted for the sequential introduction of rare codon rich regions (RCRRs) through site-directed mutagenesis (see Appendix B). The SDMs targeting site 2 and 9 did not yield positive clones, despite several attempts in which PCR conditions and template concentrations were varied. The remaining 7 mutations were introduced successfully as confirmed by sequencing (Fig. 9).

The overexpression strain BL21 (DE3) was transformed with the six positive plasmid constructs and protein expression was induced with IPTG. The results of the protein overexpression analysis, based on immuno-detection with an anti-flag antibody, are shown in figure 10.

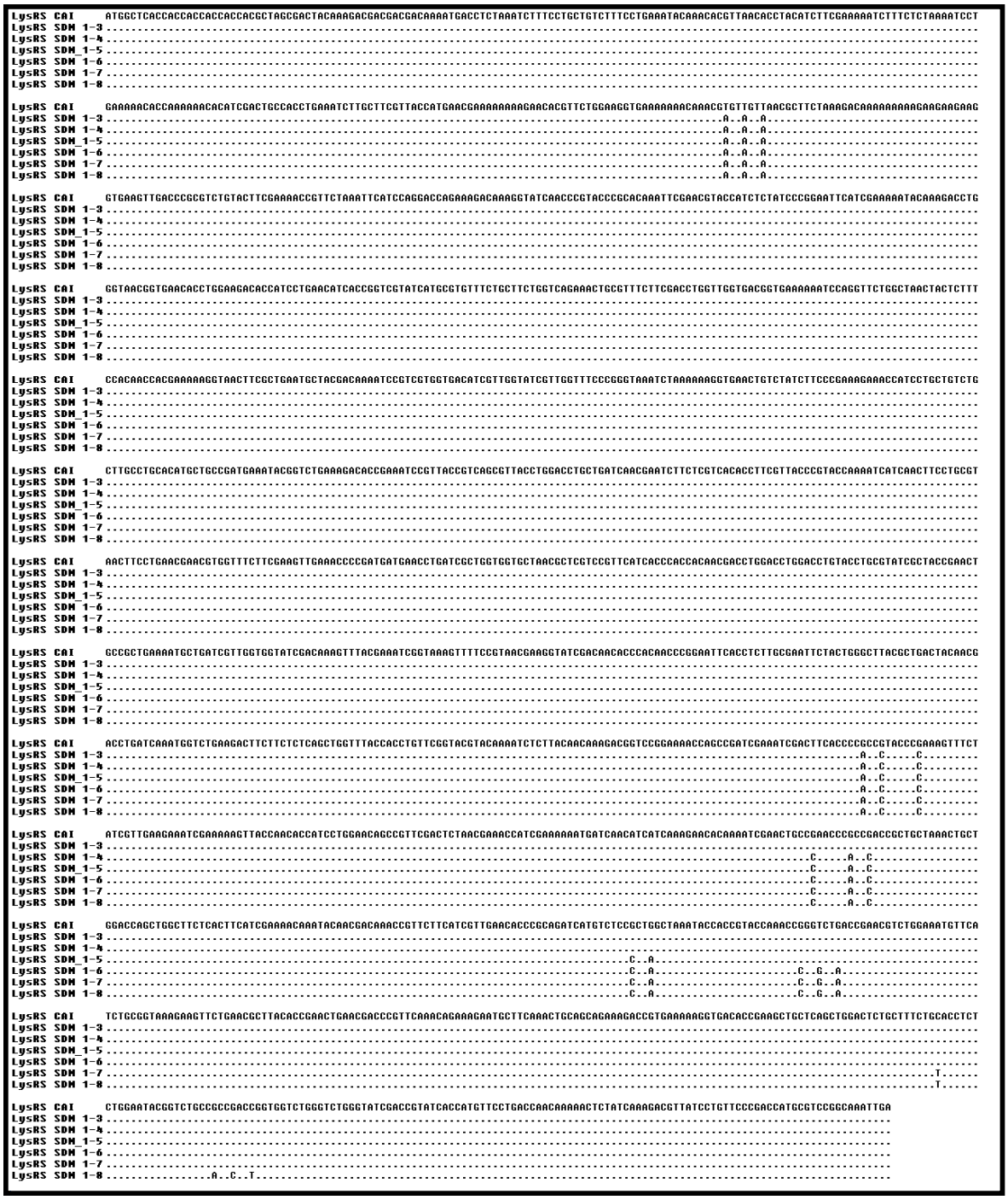


Figure 9. Sequence alignment between the CAI-optimised *Pf* LysRS and six synonymous genes, showing the sequentially increasing number of RCRRs from LysRS SDM 1-3 to LysRS SDM 1-8.

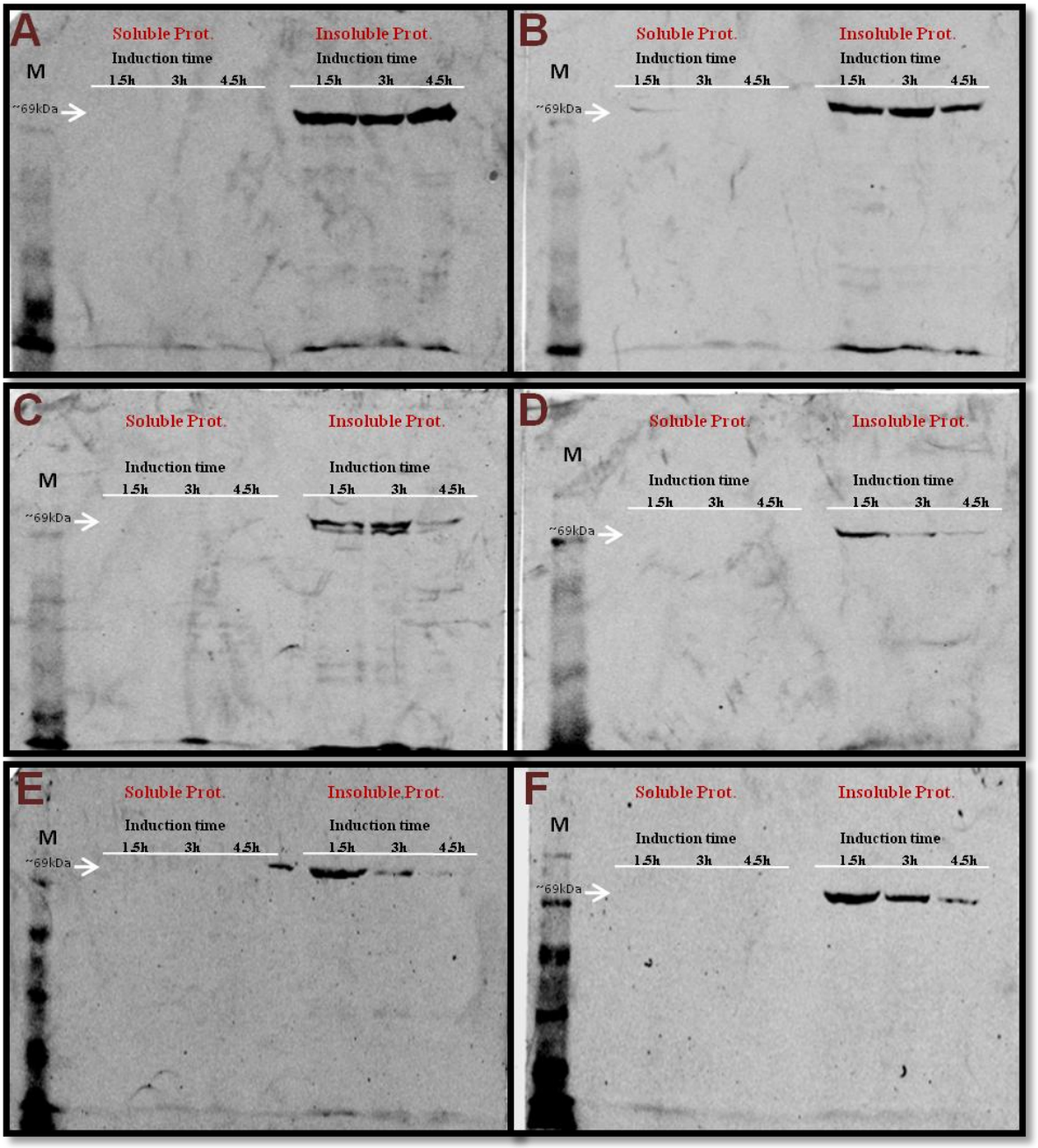


Figure 10. Western blots of soluble and insoluble protein fractions of six *Pf* LysRSs after different times of induction. The proteins originated from the overexpression of synonymous LysRS genes with increasing numbers of RCRRs : A) SDM 1-3, B) SDM 1-4, C) SDM 1-5, D) SDM 1-6, E) SDM 1-7, F) SDM 1-8. The Flag-tagged LysRSs were immuno-detected, using a monoclonal anti-Flag primary antibody and a fluorescently labeled secondary antibody. Each lane contained 50 µg of total protein. M = prestained 7-175 kDa protein ladder.

All constructs resulted in the expression of the LysRS with the correct molecular weight of ~69 kDa. However, the introduced RCRRs from site 1 to 8, did not improve protein solubility. In fact, with the exception of a faint band in SDM 1-4 after the shortest induction time (Fig. 10: B), no soluble LysRS was detected. Similarly, the amount of overexpressed protein in the insoluble fractions also generally decreased along with the induction time. This was most obvious in the LysRS bands, of SDM 1-6, 1-7 and 1-8 (Fig. 10: D, E, F). Further, in the insoluble protein fractions of SDM 1-3, 1-4 and 1-5 (Fig. 10: A, B, C) many additional bands of smaller molecular weight and weak signal intensity were detected and in the insoluble fraction of SDM 1-5 a more intense additional band of slightly smaller molecular weight than the LysRS was visible during the entire induction period.

Effect of gene harmonisation:

The harmonised gene was successfully integrated in the pET19b vector and sequencing confirmed its identity (Sequence in Appendix B; Comparison of RSCU between native and harmonised gene in Appendix C). An overexpression analysis of this construct showed that gene harmonisation did not have the desired effect of improving solubility of the *Pf* LysRS. On the contrary, the resulting protein was only found in the insoluble fraction and, overall, protein expression was lower than in the CAI-optimised gene and the other synonymous genes enriched in RCRRs (Fig. 11). Additional to the expected band of ~69 kDa, another band with a molecular weight of ~22 kDa was detected at all three time points of induction. This particular band, which is not very visible on the single gel shown in figure 3, was also detected at a later stage of this work (see Fig. 17: B and Fig. 18; after extraction and purification of the harmonised LysRS, respectively).

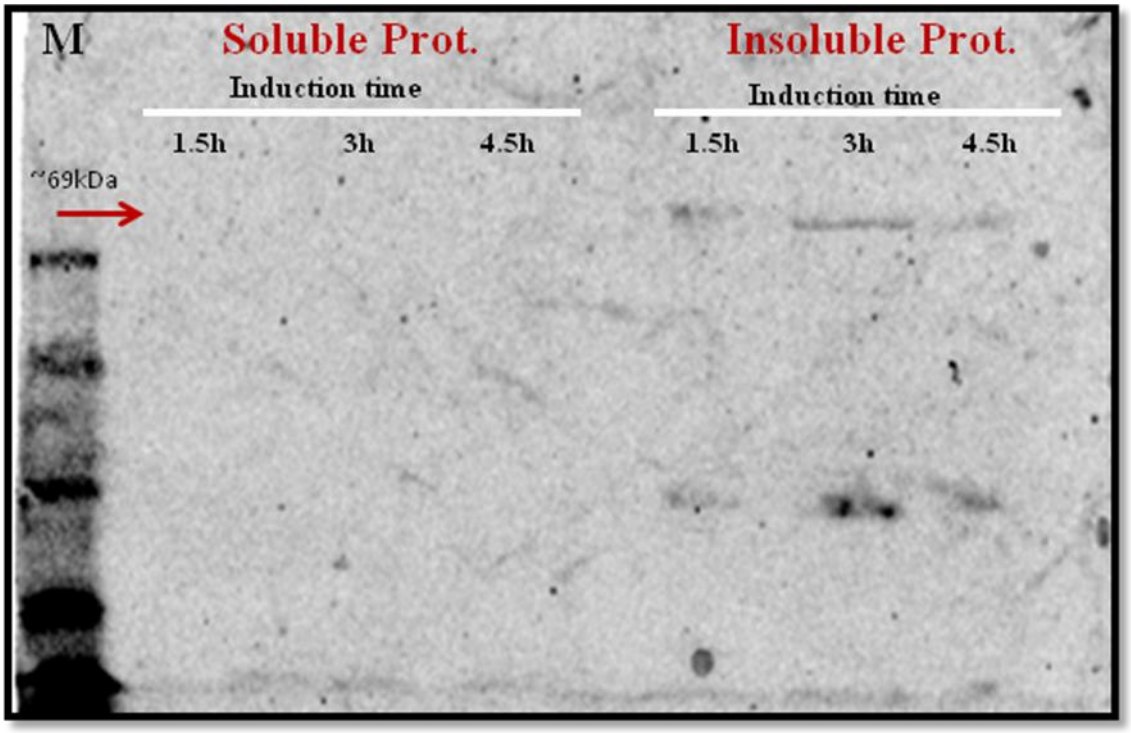


Figure 11. Western blot of soluble and insoluble fractions of the *Pf* LysRS, after different times of induction. The proteins originated from overexpressing the harmonised LysRS gene. The Flag-tagged LysRS was immuno-detected, using a monoclonal anti-Flag primary antibody and a fluorescently labeled secondary antibody.. Each lane contained 50 µg of total protein. M = prestained 7-175 kDa protein ladder.

4.2 The ability of differently optimised synonymous LysRS genes in rescuing a temperature sensitive *E. coli* strain

To test the functionality of the different *Pf* LysRS genes, a temperature rescue assay with the *E. coli* strain PALΔSΔUTR was performed. The only functional copy of the LysRS gene this strain possesses is on a temperature sensitive pMAK705 plasmid. The strain was transformed with a vector control (pET19b plasmid without *Pf* LysRS gene) or one of 5 vectors, each containing a different LysRS gene: WT or one of four *Pf* LysRS transformants (CAI-optimised, context-optimised, CAI/CON-optimised or codon harmonised) (Sequences are shown in Appendix B).

The different pMAK705 clones were grown with and without induction, at 30 ºC and 42 ºC. Unexpectedly, all pMAK705 transformants grew better at 42 ºC than at 30 ºC and the induction with IPTG had no obvious effect on growth (Fig. 12), meaning that the rescue assay could not be performed as planned.

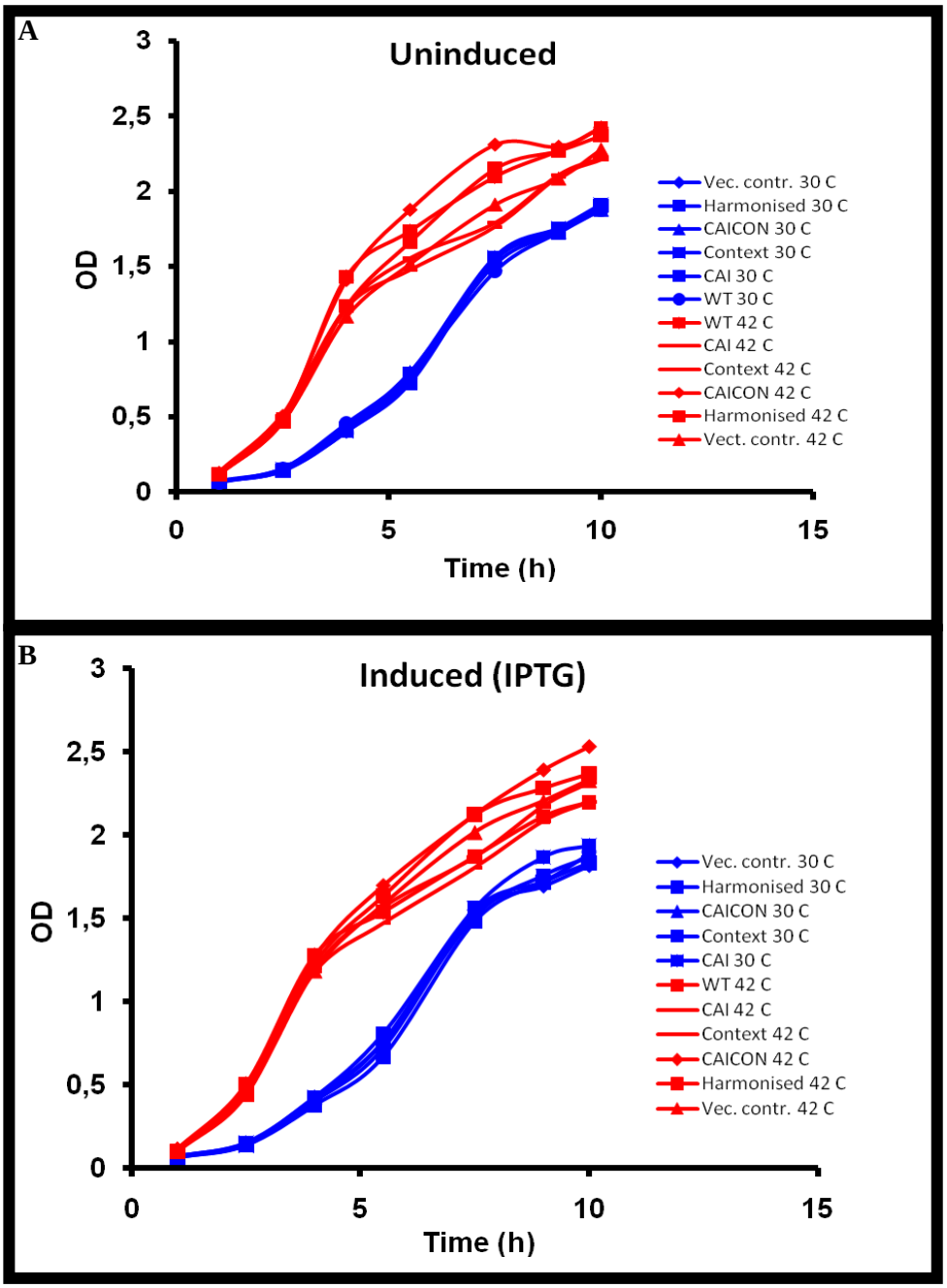


Figure 12. Growth curves at 30 °C and 42 °C of the *E. coli* strain PALΔSΔUTR transformed with either a plasmid control or one of five *Pf* LysRS gene constructs: A) without induction; B) induced with IPTG.

4.3 Codon context effect on amino acid misincorporation assessed by a heat stability β-gal assay

By means of a strain/vector system that permitted restoration of β–gal activity through α-complementation, the effects of a context de-optimised lacZα gene on the heat stability of the β–gal was tested. Samples were collected at different inactivation time points and heat stability was measured, using ONPG as substrate and a colorimetric measurement of the product o-nitrophenol (Fig. 13).

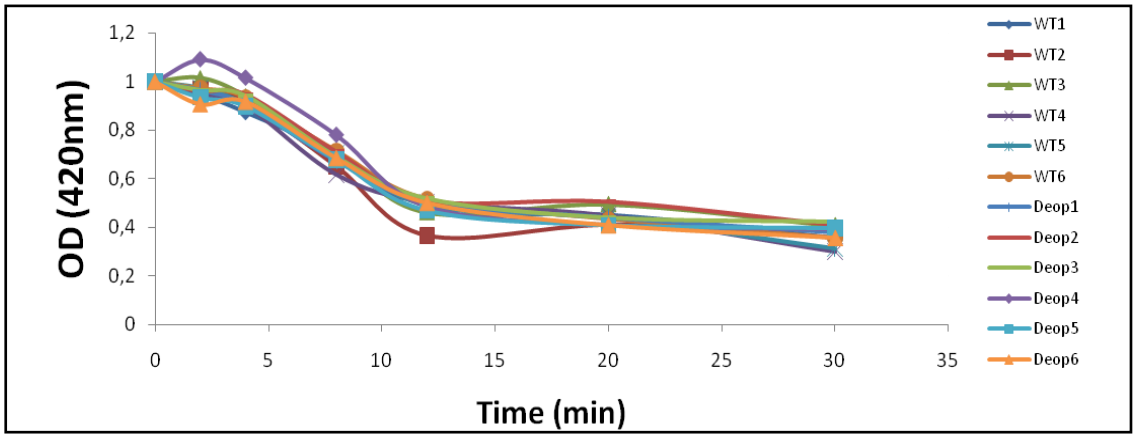


Figure 13. β-gal assay with KRX cells expressing a WT or context-de-optimised gene for the α-peptide. Shown is the OD_{420nm}, a measure for the amount of o-nitrophenol and thus β-gal activity, versus time of heat inactivation. The enzymatic activity of the β-galactosidase decreased with the length of heat inactivation but no differences were observed between the WT and context de-optimised lacZα gene. For both the WT and the mutant gene six biological replicates were analysed.

.

As expected, the activity of the β-galactosidase decreased with the length of heat inactivation; the strongest decrease of enzymatic activity being observed during the first 12 minutes. However, differences in heat stability between the gene products of the WT and context de-optimised lacZα gene were not observed.

4.4 Sample preparation for mass spectrometry analysis

In order to gain more specific information on the site specific misincorporation of amino acids, the His-tagged LysRS and lacZα transformants were purified and prepared for mass spectrometry.

Purification of the lacZα gene product (α-peptide of β-galactosidase)

In preparation for the mass spectrometric detection of aa misincorporation due to bad codon context, total protein from KRX cells was extracted (soluble and insoluble fractions) and analysed by western blotting to assess the amount and quality of overexpressed protein. The coomassie staining and immuno-detection of the extracts are shown in figures 14 and 15, respectively.

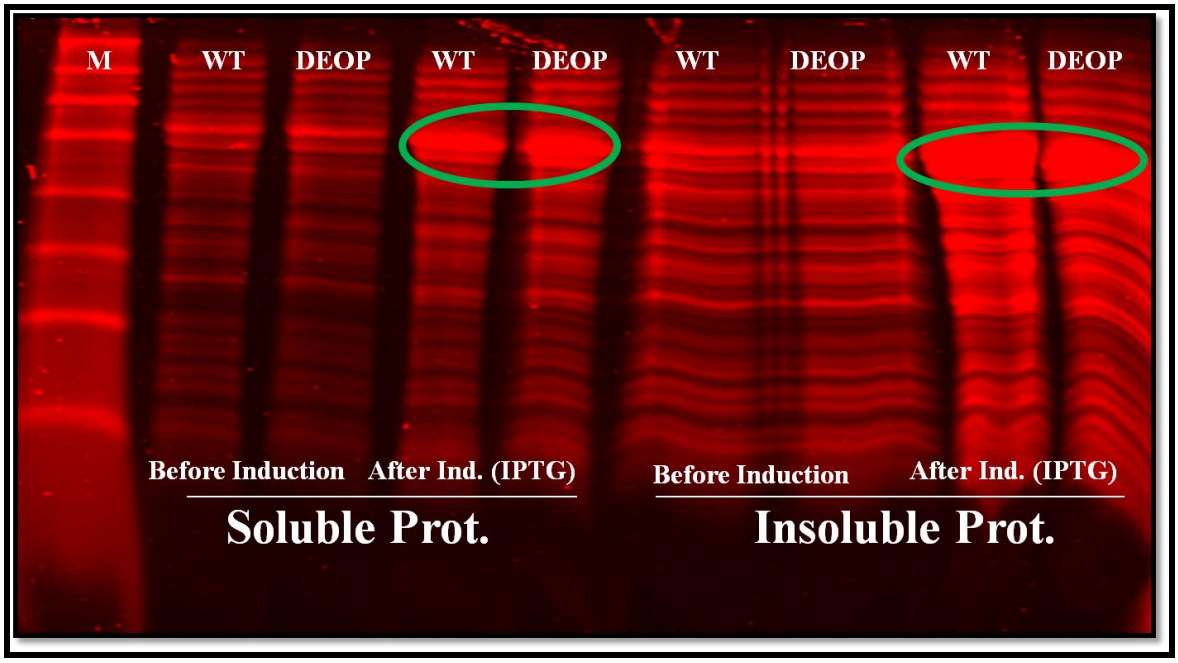


Figure 14. Coomassie staining of the soluble and insoluble protein fractions of KRX cells, transformed with the β–gal α-peptide, before and after IPTG induction. Each lane contained 100 µg of total protein. Size standard = prestained 7-175 kDa protein ladder.

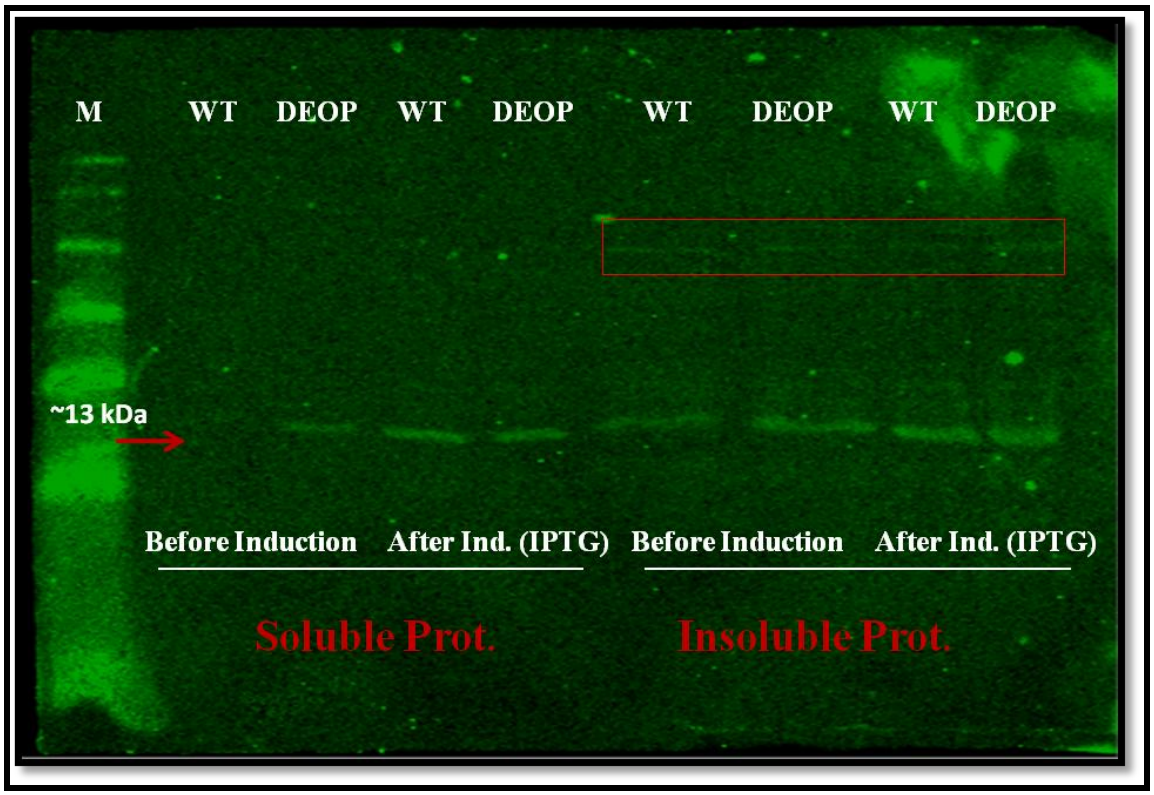


Figure 15. Western Blot of the soluble and insoluble protein fractions of KRX cells, transformed with the β–gal α-peptide, before and after IPTG induction. The His-tagged α-peptide was detected using immuno-detection with an anti-His antibody. Each lane contained 100 µg of total protein. Size standard = prestained 7-175 kDa protein ladder.

The coomassie staining showed a clear increase in the amount of a specific protein after induction, both in the soluble and insoluble fractions of WT and de-optimised samples. However, the molecular weight of the induced protein (marked with green circles in fig. 14) was much higher than the expected one. Despite the unexpected molecular weight of the seemeingly induced protein, the samples were purified using the anti-His IMAC resin. The following immuno-detection then resulted in the visualisation of bands corresponding to the expected molecular weight of ~13 kDa of the his-tagged α-peptide (Fig. 15). However, the signals were only visible after increasing the secondary antibody concentration from 1:1000 to 1:500, indicating that only small amounts of protein were present. Weak bands, corresponding to the higher molecular weight bands that were dominant in the coomassie staining were also identified by imuno-detection in the insoluble fraction (framed red in figure 15). Since the amount of protein was not sufficient for mass spectrometry analysis of aa misincorporation, thus far no further steps were taken to study the α-peptide.

Purification of the LysRS

The approach that was employed for the α-peptide was also used for the *Pf* LysRS gene transformants: WT, CAI-optimised, Context-optimised, CAI/CON-optimised, Harmonised and hybrid gene 2 - DIG 2 (construct specifications in Appendix B). Except the harmonised gene, all other constructs were the result of previous work in the lab [50].

Following the protein extraction and in line with previous experiments, the coomassie staining and the immuno-detection with anti-his antibody showed that solubility was not significantly improved by using any of the optimisation approaches (Fig. 16 and 17, respectively).

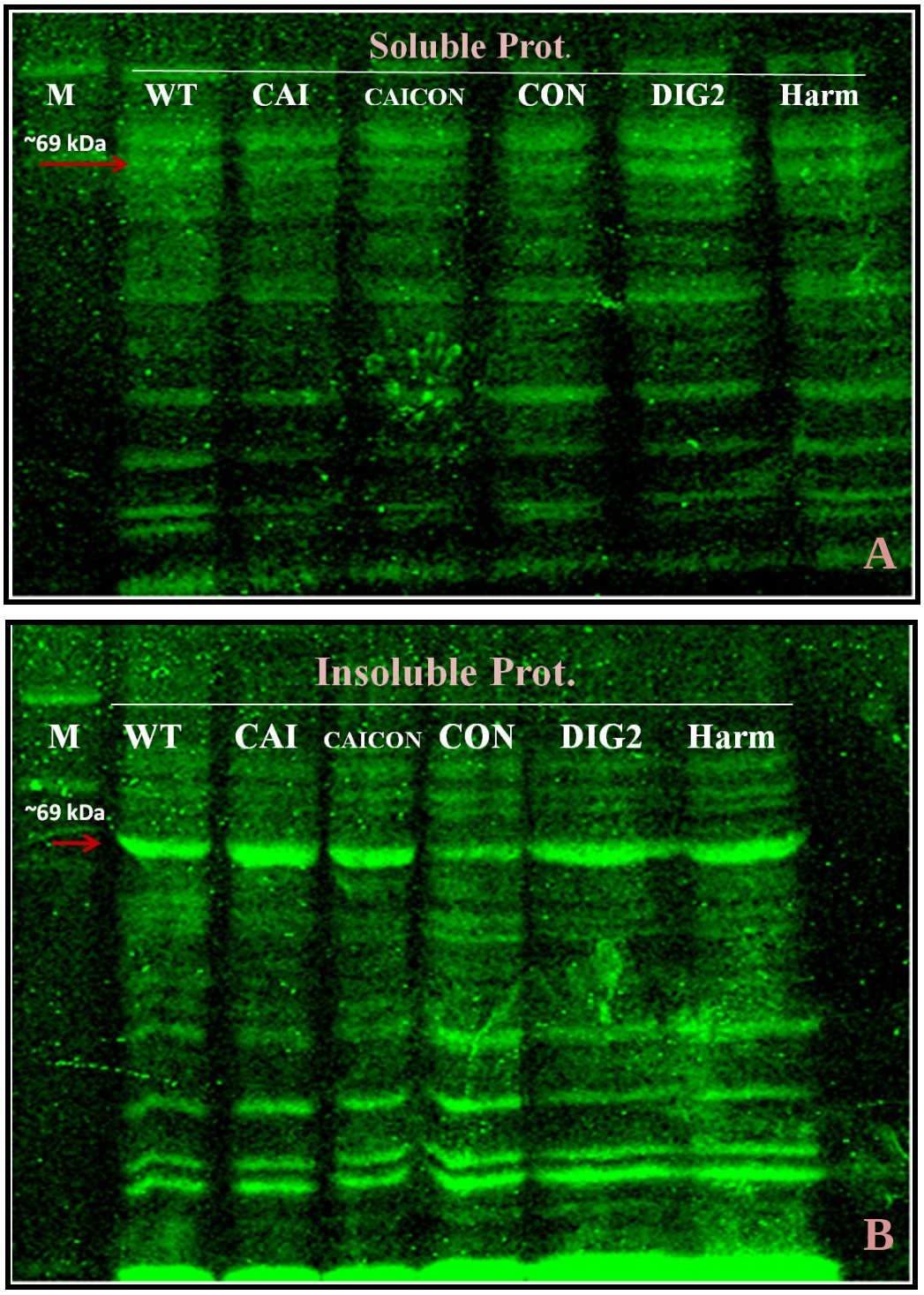


Figure 16. Coomassie staining of A) Soluble fractions and B) Insoluble fractions of: WT and 5 different *Pf* LysRS transformants: CAI-optimised, CAICON-optimised, Context-optimised, Hybrid gene 2 (Dig2) and Harmonised. Each lane contained 100 µg of total protein. Size standard = prestained 7-175 kDa protein ladder.

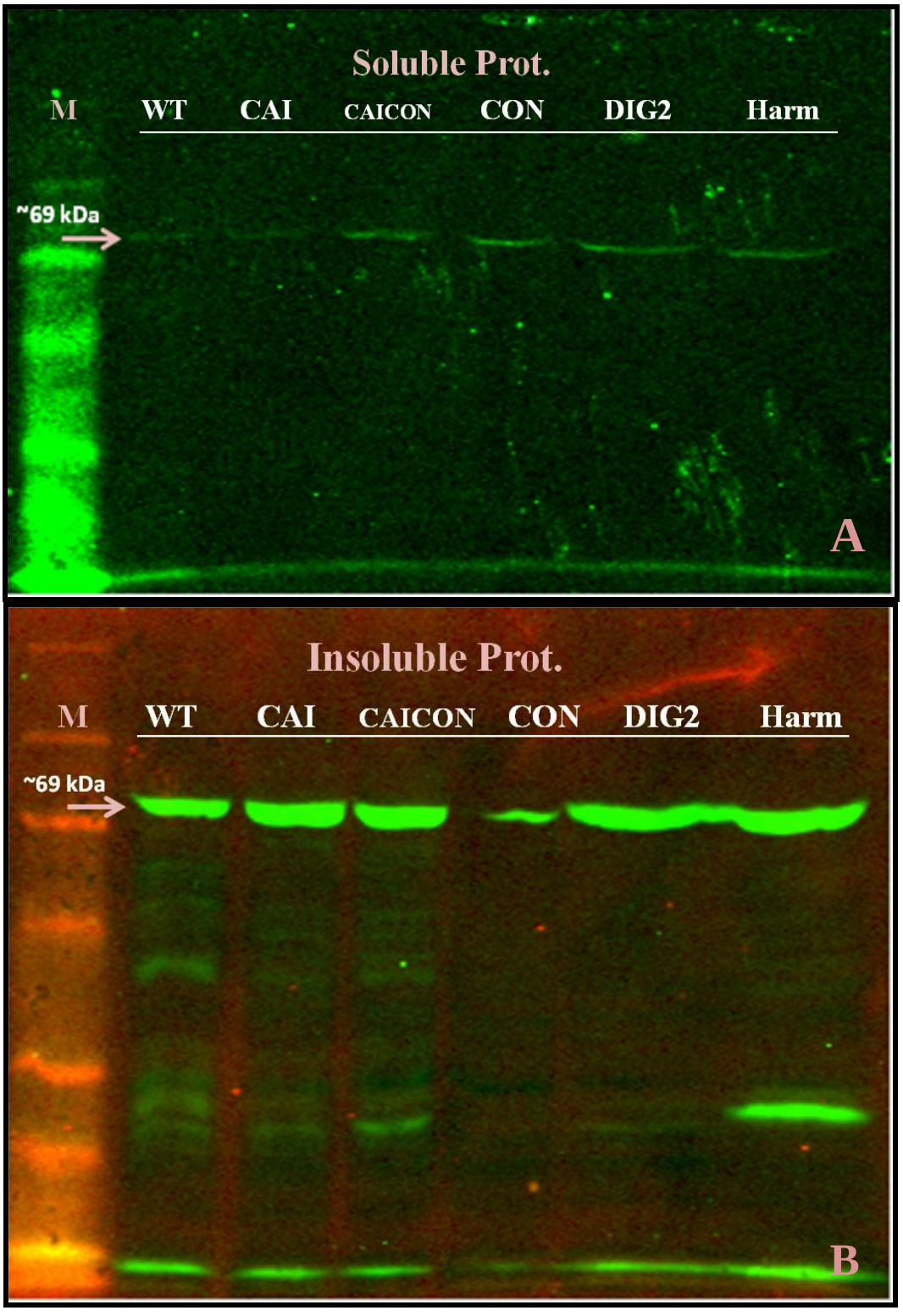


Figure 17. Western Blot immuno-detection. A) Soluble fractions and B) Insoluble fractions of WT and five different *Pf* LysRS transformants: CAI-optimised, CAICON-optimised, Context-optimised, Hybrid gene 2 (Dig 2) and Harmonised. The Flag-tagged LysRS bands were detected using a anti-flag antibody 1:1000. Each lane contained 100 µg of total protein. Size standard = prestained 7-175 kDa protein ladder

After Western blotting and Immuno-staining, intense bands of insoluble fractions could be observed for all the LysRS transformants, except for codon-context optimisation, which was already expected to cause protein suppression, based on previous results. Also some low molecular weight protein bands were observed (Fig. 17: B), which were most abundant in the WT and virtually missing in the DIG 2 extracts. Since sufficient amounts of target protein were present in the insoluble fractions, the following purification step was performed for all insoluble fractions, except the context-optimised transformant. This sample was replaced by the DIG 2 (context-optimised gene with a different his-tag).

The purification results of the chosen samples indicated that a substantial amount of LysRS protein was successfully obtained in all of them, which is visible by the intense bands on the coomassie stained gels (Fig. 18). By analysing the gel, the WT sample showed again a higher number of additional low molecular bands compared with all other samples and DIG 2 did not show any additional bands. It was detected that each low molecular weight band identity can be visualised at the same level in different samples. That is easily visible comparing the samples WT, CAI and CAICON (Fig. 17). Three of these bands were selected (indicated in fig. 18) and prepared for mass spectrometry analysis alongside the five full-length LysRS samples. Also an interesting target for mass spectrometry would have been a single, well-defined low molecular weight band in the harmonised construct sample, but amounts of protein were too low to move on the mass spectrometry analysis just yet.

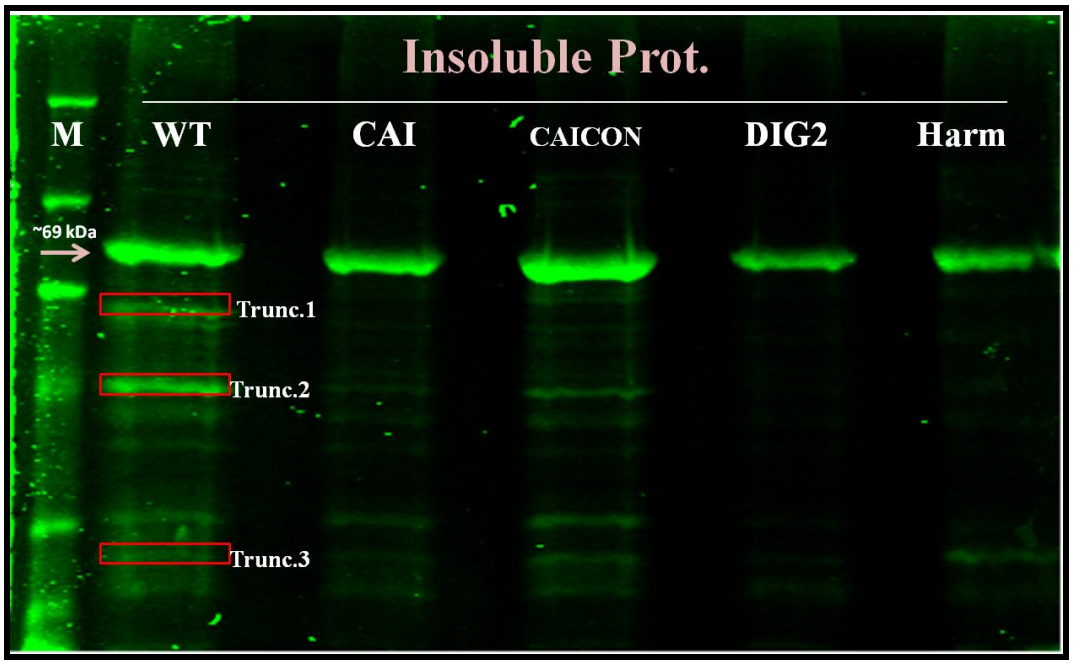


Figure 18. Coomassie staining after His-tag purification of insoluble protein fractions originating from the overexpression of the WT and 4 synonymous *Pf* LysRS genes: CAI-optimised, CAICON-optimised, Hybrid gene 2 (Dig2) and Harmonised. Additionally to the main bands of ~69 kDa, the 3 truncated proteins indicated in the figure were also selected for mass spectrometry analysis. Each lane contained 10 μg of total protein. Size standard = prestained 7-175 kDa protein ladder.

5 Discussion

5.1 Effects of RCRRs and codon harmonisation on protein expression and protein solubility of the plasmodial LysRS

Effect of RCRRs

The translation machinery of *P. falciparum* remains largely uncharacterized and particularly the location of ARS (Aminoacyl-tRNA synthetases) and the tRNA isoacceptors abundance in *Plasmodium* are still unclear [52-53]. In *E. coli*, like in many other organisms, a close relationship exists between tRNA copy number and codon usage patterns [54], but *P. falciparum* has a peculiar genome, containing only one copy of each tRNA isoacceptor gene [55]. This fact might indicate that this organism controls its translation rate in a distinct way, not yet known. Furthermore, the lack of tRNA abundance data for this organism made it necessary to define rare codons purely based on codon usage, which means that some of the codons that were defined as "rare" may in fact be read by abundant tRNAs in *P. falciparum*. This means that some of the supposed functionally relevant RCRRs were probably not identified correctly and thus their optimisation for *E. coli* did not contribute to the desired solubility improvement effect. Considering these points, it is not entirely surprising that the overexpression analysis of soluble and insoluble fractions of the rare codon-enriched LysRS genes showed no improvement in protein solubility.

The different bands of smaller molecular weight shown in figure 10: A, B and C could be an indication for degraded or truncated proteins, which suggests that the protein is being degraded due to a control mechanism of the heterologous cell or that the translation process could have been interrupted at specific sites of the ORF on that cell. One possible reason is that the introduction of RCRRs may have caused a prolonged pause in the ribosomal traffic at specific sites, leading to a premature dissociation of peptidyl-tRNAs from ribosomes (drop-off) [56] and resulting in the production of several truncated forms of the full-length LysRS of ~69 kDa.

Since the respective low molecular weight bands were detectable with the anti-Flag antibody, they must have been truncated on the C-terminal end (the end opposite to the Fag-tag). To obtain an indication of potential problematic positions, it would be important to analyze the mRNA secondary structure of the different SDM samples. It is possible that strong mRNA secondary structures, such as hairpins, were disturbing the translation process by causing ribosomal drop-off or frame-shifts and in the end, that information could be correlated with the upcoming data from mass spectrometry analysis.

A remarkable result was the apparent decrease of LysRS expression with increasing times of induction. An inducer like IPTG strongly intensifies the target transcript production, and consequently a higher number of ribosomes translates the induced mRNA. If tRNA isoacceptor abundance for the induced protein is a limiting factor in the cell, induction could lead to significant changes in tRNA abundance and could result in ribosomal traffic jams and therefore, to protein aggregation and ultimately protein degradation. Nevertheless, analysing the gels (Fig. 10), there was no evident increase in the amount of truncated or degraded proteins along the induction time (the number of small molecular weight bands was similar comparing with the first time point). Another explanation for the seemingly reduction of overexpressed protein over the course of induction could be that high amounts of the insoluble fractions of the LysRS were toxic for the host cell and therefore, activated a downregulation system to decrease expression of the toxic protein. In this case, and if the host keeps growing and producing other proteins, it could be expected that with longer induction time, the overexpressed protein would represent a smaller and smaller fraction of the total protein.

The fraction before induction (time point zero) was not collected, but pervious results [50] showed that a soluble form of the LysRS was usually more abundant before induction. This previous observation and the current finding that in SDM 1-4 there was a small amount of soluble protein present after 1,5h but not after longer induction times (figure 10: B), suggests that uninduced expression of the LysRS may have translation dynamics that are more beneficial for the folding of this enzyme in *E. coli*.

Effect of gene harmonisation

The rebuilding of the RSCU pattern of the *Pf* LysRS gene in the heterologous host *E. coli*, i.e. the codon harmonisation of the *Pf* LysRS gene, did not result in improved LysRS solubility (Fig. 11). One possible reason for the obtained results in this optimisation approach, and identically to the previous study, is that synonymous codon substitutions might be having a strong influence in the secondary structure of the mRNA chain.

As well as already mentioned for the RCRRs introduction, concentration adjustment of intracellular tRNA isoacceptor molecules is needed to solve codon usage disparities. That could be planned in a very limited way by using specific plasmids encoding rare tRNAs for the heterologous system [57] and consequently avoiding their scarcity during the translation process.

5.2 The ability of differently optimised synonymous LysRS genes in rescuing a temperature sensitive *E. coli* strain

Unfortunately, the temperature rescue assay did not work as expected and experiments had to be aborted after a few trials. Most likely, the observed lack of temperature sensitivity was due to a confusion of strain-ID in a collaborating lab. Other possible reasons for the unexpected results could be that the genomic LysRS of the used strain regained its functionality through random mutations or that, through a re-integration event of the plasmidic copy of the LysRS gene back into the genome, the synthetase gene was no longer under the control of the temperature-sensitive plasmidic induction system.

5.3 Codon context effects on mistranslation: assessed by a heat stability β-gal assay

One aim of this study was to assess the potential of a novel reporter system to test codon context effects on translation accuracy and efficiency *in vivo*. As reporter, the *E. coli* enzyme β–galactosidase (β-gal) was chosen (for an active quaternary structure of the enzyme see Appendix D). Since the α-peptide establishes an important dimer–dimer interaction in the functional β-gal [58], the assay design was based on the hypothesis

that, if context de-optimisation of the lacZα gene increases mistranslation, then the heat stability of the α-complemented, functional β–gal enzyme should be measurably decreased.

In this study, the substitutions introduced in the lacZα gene (Codon context de-optimisation in Appendix C) did not show an effect on the heat stability of the expressed protein. One reason for this could be that the targeted lacZα gene is relatively short compared to the entire β-galactosidase coding sequence and therefore, the number of aa misincorporations was potentially not sufficient to cause a pronounced heat stability difference between samples. Further, based on previous bioinformatics analysis codon context-related mistranslation is likely to result in the misincorporation of chemically similar amino acids [59], which could make functional assays relatively insensitive to mistranslation. Thus, the usefulness of functional assays to detect mistranslation is probably strongly dependent on choosing a very sensitive model protein. The β-galactosidase may still prove to be a good target for this purpose, but further tests are necessary to assess this potential. More immediately, mass spectrometry analysis could give important answers to the question of whether the observed lack of differences in heat stability were actually due to low rates of mistranslation, the misincorporation of chemically similar amino acids or the lack of sensitivity in the observed model system.

5.4 Samples preparation for mass spectrometry analysis

Purification of the lacZα

The induction with IPTG clearly increased the expression of a high molecular weight protein (~46 kDa) in both the soluble and insoluble fraction of the total protein extracts (encircled in fig. 14). Since the expected molecular weight of the α-peptide is much lower (~13 kDa), the appearance of these bands is difficult to explain. One explanation could be that the observed band may have consisted of the α-peptide associated to some fraction of the genome-encoded LacZΩ gene product, but the molecular weight of the two monomeric peptides combined would already exceed the observed size. Further, since the proteins were separated by SDS-PAGE, any quarternary protein structures should have been denatured.

In contrast to the results obtained by coomassie staining, the immuno-detection identified weak but clear bands corresponding to the size of the α-peptide, giving a good indication that at least a small amount of the α-peptide was present in the samples. However, since the amount of identifiable α-peptide was not sufficient for mass spectrometry analysis of aa misincorporation, thus far no further steps were taken to prepare the α-peptide for mass spectrometry.

Purification of the LysRS

In agreement with previous results [50], in most cases a high yield of LysRS was obtained in the insoluble protein extracts but little or no LysRS was present in a soluble form. Further, lower molecular weight bands were identified in almost all samples, but synonymous LysRS genes differed repeatedly in the amount of these lower molecular weight proteins (Fig. 17: B and fig. 18). These bands could have been related to the production of truncated LysRS or the degradation of insoluble LysRS (Fig. 17). Although there was coherence between the results, before and after purification, a positive control could have been useful to allow a better comparison of band intensities between gels.

It would have been interesting to analyse the mRNA secondary structure of all the mutant genes in order to understand the possible causes for the production of truncated or degraded proteins. The amount of purified protein for the WT, CAI, CAICON, DIG2 and Harmonised LysRS transformants was sufficient to proceed to mass spectrometry, but technical problems caused a delay in analysing these samples.

6 Conclusions / Future work

Translation is a very complex process that, in addition to host-specific variables and environmental conditions [2], depends on multiple other factors. Because of this complexity, gene design for heterologous expression remains a challenging task and robust rules that establish clear relationships between the different variables are difficult to extract.

In this work, the effect of several mRNA primary structure features, such as codon usage, RCRRs and codon context were studied. Although it was not possible within the relatively short time of the project to extract specific gene design rules for the optimisation of *P. falciparum* genes for heterologous gene expression, some interesting avenues for further research were opened. The first gene optimisation approaches used in this work: the sequential introduction of RCRRs in the CAI optimised *Pf* LysRS gene as well as its harmonisation, did not show the expected improvement on protein solubility. After extraction of the soluble and insoluble protein fractions, some low molecular weight bands were identified, possibly corresponding to truncated or degraded LysRS proteins. This suggests that a ribosomal drop-off effect may have occurred due to the presence of strong secondary structures along the mRNA chain or ribosomal traffic jams. Since *P. falciparum* has a peculiar genome and the abundance of the complete pool of cellular tRNA isoacceptors is not known, this could be a limiting for protein heterologous expression and therefore, the gene harmonisation and the putative RCRRs introduction probably do not play the expected function in the heterologous translation dynamic. Moreover, the synonymous codon substitutions established for both optimisation approaches might also have a strong influence on the mRNA secondary structure, disturbing the translation system.

Another interesting result was the decrease of LysRS along induction time and the particular identification of a thinner soluble fraction in the transformant SDM 1-4, for the first induction time. These observations might indicate that the overexpression levels reached during induction are so strong that tRNA limitation becomes a relevant factor that in turn changes translation dynamics and potentially causes ribosomal traffic jams. Uninduced expression of the LysRS might reduce these effects and may actually be a way forward in expressing challenging plasmodial proteins.

Future work should include a more detailed mRNA secondary structure analysis as well as the quantification of the complete pool of cellular tRNAs in *P. falciparum* to better understand its translation mechanism and to adjust the respective tRNA concentrations in the heterologous host. The upcoming mass spectrometry analysis of the *Pf* LysRS may give important clues to what extent different gene optimisation strategies affect mistranslation and ribosomal drop-off and the identification of the specific aa sites that are involved in these potential translation problems may point towards specific codons and codon combinations, worth investigating in more detail.

Finally, the codon context-de-optimisation of the LacZα gene did not show significant heat stability differences between the native a context-de-optimised transformants through the β-gal assay. It is possible that the synonymous substitutions introduced in this gene were not sufficient to obtain a significant effect on protein heat stability, or the performed assay was not sensitive enough to detect significant differences between the native and codon context-de-optimised transformants.

Mass spectrometry analysis may also give insight into the effect of context on the misincorporation of aa, but first the encountered overexpression problems will have to be solved, so that sufficient amounts of α-peptide become available for further analysis.

References

1. Elliott, W. and D. Elliott, *Biochemistry and Molecular Biology*. Third Edition ed. 2005, Oxford: University Press.
2. Welch, M., A. Villalobos, C. Gustafsson, and J. Minshull, *You're one in a googol: optimizing genes for protein expression.* Journal of the Royal Society Interface, 2009. **6**: p. S467-76.
3. Kozak, M., *Regulation of translation via mRNA structure in prokaryotes and eukaryotes.* Gene, 2005. **361**: p. 13-37.
4. Gebauer, F. and M.W. Hentze, *Molecular mechanisms of translational control.* Nature Reviews Molecular Cell Biology, 2004. **5**(10): p. 827-835.
5. Plotkin, J. and G. Kudla, *Synonymous but not the same: the causes and consequences of codon bias*, in *Nature Reviews Genetics* 2011. p. 32-42.
6. Cannarrozzi, G., N.N. Schraudolph, M. Faty, P. von Rohr, M.T. Friberg, A.C. Roth, P. Gonnet, G. Gonnet, and Y. Barral, *A Role for Codon Order in Translation Dynamics.* Cell, 2010. **141**(2): p. 355-367.
7. Zhang, W.C., W.H. Xiao, H.M. Wei, J. Zhang, and Z.G. Tian, *mRNA secondary structure at start AUG codon is a key limiting factor for human protein expression in Escherichia coli.* Biochemical and Biophysical Research Communications, 2006. **349**(1): p. 69-78.
8. Wang, F.P. and H. Li, *Codon-pair usage and genome evolution.* Gene, 2009. **433**(1-2): p. 8-15.
9. Salim, H.M.W. and A.R.O. Cavalcanti, *Factors influencing codon usage bias in genomes.* Journal of the Brazilian Chemical Society, 2008. **19**(2): p. 257-262.
10. Grantham, R., C. Gautier, M. Gouy, M. Jacobzone, and R. Mercier, *Codon Catalog Usage Is a Genome Strategy Modulated for Gene Expressivity.* Nucleic Acids Research, 1981. **9**(1): p. R43-R74.
11. Tats, A., T. Tenson, and M. Remm, *Preferred and avoided codon pairs in three domains of life.* Bmc Genomics, 2008. **9**: p. 463.
12. Hershberg, R. and D. Petrov, *Selection on Codon Bias.* Annual Review of Genetics, 2008. **42**.
13. Sharp, P.M. and W.H. Li, *The Codon Adaptation Index - a Measure of Directional Synonymous Codon Usage Bias, and Its Potential Applications.* Nucleic Acids Research, 1987. **15**(3): p. 1281-1295.
14. Ikemura, T., *Correlation between the Abundance of Escherichia-Coli Transfer-RNAs and the Occurrence of the Respective Codons in Its Protein Genes.* Journal of Molecular Biology, 1981. **146**(1): p. 1-21.
15. Bennetzen, J.L. and B.D. Hall, *Codon Selection in Yeast.* Journal of Biological Chemistry, 1982. **257**(6): p. 3026-3031.
16. Ikemura, T., *Codon Usage and Transfer-RNA Content in Unicellular and Multicellular Organisms.* Molecular Biology and Evolution, 1985. **2**(1): p. 13-34.
17. Horn, D., *Codon usage suggests that translational selection has a major impact on protein expression in trypanosomatids.* Bmc Genomics, 2008. **9**: p. 2.
18. Irwin, B., J.D. Heck, and G.W. Hatfield, *Codon Pair Utilization Biases Influence Translational Elongation Step Times.* Journal of Biological Chemistry, 1995. **270**(39): p. 22801-22806.
19. Widmann, M., M. Clairo, J. Dippon, and J. Pleiss, *Analysis of the distribution of functionally relevant rare codons.* Bmc Genomics, 2008. **9**: p. 207.

20. Zhang, G., M. Hubalewska, and Z. Ignatova, *Transient ribosomal attenuation coordinates protein synthesis and co-translational folding.* Nature Structural & Molecular Biology, 2009. **16**(3): p. 274-280.
21. Komar, A.A., *A pause for thought along the co-translational folding pathway.* Trends in Biochemical Sciences, 2009. **34**(1): p. 16-24.
22. Tuller, T., A. Carmi, K. Vestsigian, S. Navon, Y. Dorfan, J. Zaborske, T. Pan, O. Dahan, I. Furman, and Y. Pilpel, *An Evolutionarily Conserved Mechanism for Controlling the Efficiency of Protein Translation.* Cell, 2010. **141**(2): p. 344-354.
23. Moura, G., M. Pinheiro, J. Arrais, A.C. Gomes, L. Carreto, A. Freitas, J.L. Oliveira, and M.A.S. Santos, *Large Scale Comparative Codon-Pair Context Analysis Unveils General Rules that Fine-Tune Evolution of mRNA Primary Structure.* Plos One, 2007. **2**(9): p. e847.
24. Moura, G., M. Pinheiro, R. Silva, I. Miranda, V. Afreixo, G. Dias, A. Freitas, J.L. Oliveira, and M.A.S. Santos, *Comparative context analysis of codon pairs on an ORFeome scale.* Genome Biology, 2005. **6**(3): p. R28.
25. Boycheva, S., G. Chkodrov, and I. Ivanov, *Codon pairs in the genome of Escherichia coli.* Bioinformatics, 2003. **19**(8): p. 987-998.
26. Moura, G., M. Pinheiro, J. Arrais, A.C. Gomes, L. Carreto, A. Freitas, J.L. Oliveira, and M.A.S. Santos, *Large Scale Comparative Codon-Pair Context Analysis Unveils General Rules that Fine-Tune Evolution of mRNA Primary Structure.* Plos One, 2007. **2**(9): p. -.
27. Moura, G.R., J.P. Lousado, M. Pinheiro, L. Carreto, R.M. Silva, J.L. Oliveira, and M.A.S. Santos, *Codon-triplet context unveils unique features of the Candida albicans protein coding genome.* Bmc Genomics, 2007. **8**: p. 444.
28. Mahlen, S.D., S.S. Morrow, B. Abdalhamid, and N.D. Hanson, *Analyses of ampC gene expression in Serratia marcescens reveal new regulatory properties.* Journal of Antimicrobial Chemotherapy, 2003. **51**(4): p. 791-802.
29. Rosenbaum, V., T. Klahn, U. Lundberg, E. Holmgren, A. Von Gabain, and D. Riesner, *Co-existing structures of an mRNA stability determinant. The 5′ region of the Escherichia coli and Serratia marcescens ompA mRNA.* Journal of Molecular Biology, 1993: p. 229.
30. Hall, M.N., J. Gabay, M. Debarbouille, and M. Schwartz, *A Role for Messenger-RNA Secondary Structure in the Control of Translation Initiation.* Nature, 1982. **295**(5850): p. 616-618.
31. Wang, L.J. and S.R. Wessler, *Role of mRNA secondary structure in translational repression of the maize transcriptional activates L-C.* Plant Physiology, 2001. **125**(3): p. 1380-1387.
32. Titov, I.I., D.G. Vorobiev, V.A. Ivanisenko, and N.A. Kolchanov, *A fast genetic algorithm for RNA secondary structure analysis.* Russian Chemical Bulletin, 2002. **51**(7): p. 1135-1144.
33. Sheppard, K., J. Yuan, M.J. Hohn, B. Jester, K.M. Devine, and D. Soll, *From one amino acid to another: tRNA-dependent amino acid biosynthesis.* Nucleic Acids Research, 2008. **36**(6): p. 1813-1825.
34. Tang, D.T.P., E.A. Glazov, S.M. McWilliam, W.C. Barris, and B.P. Dalrymple, *Analysis of the complement and molecular evolution of tRNA genes in cow.* Bmc Genomics, 2009. **10**: p. 188.
35. Yuan, J., K. Sheppard, and D. Soll, *Amino acid modifications on tRNA.* Acta Biochimica Et Biophysica Sinica, 2008. **40**(7): p. 539-553.

36. Ferrer, E., *Spotlight on targeting aminoacyl-tRNA synthetases for the treatment of fungal infections.* Drug News Perspect. , 2006: p. 19(6).
37. Kanaya, S., Y. Yamada, M. Kinouchi, Y. Kudo, and T. Ikemura, *Codon usage and tRNA genes in eukaryotes: Correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis.* Journal of Molecular Evolution, 2001. **53**(4-5): p. 290-298.
38. Zhang, G. and Z. Ignatova, *Generic Algorithm to Predict the Speed of Translational Elongation: Implications for Protein Biogenesis.* Plos One, 2009. **4**(4): p. e5036.
39. Angov, E., C.J. Hillier, R.L. Kincaid, and J.A. Lyon, *Heterologous Protein Expression Is Enhanced by Harmonizing the Codon Usage Frequencies of the Target Gene with those of the Expression Host.* Plos One, 2008. **3**(5): p. e2189.
40. Chen, Y.J. and M. Inouye, *The intramolecular chaperone-mediated protein folding.* Current Opinion in Structural Biology, 2008. **18**(6): p. 765-770.
41. Ziv, G., G. Haran, and D. Thirumalai, *Ribosome exit tunnel can entropically stabilize alpha-helices.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(52): p. 18956-18961.
42. Ben Yehezkel, T., G. Linshiz, H. Buaron, S. Kaplan, U. Shabi, and E. Shapiro, *De novo DNA synthesis using single molecule PCR.* Nucleic Acids Research, 2008. **36**(17): p. e107.
43. Fuglsang, A., *Codon optimizer: a freeware tool for codon optimization.* Protein Expression and Purification, 2003. **31**(2): p. 247-249.
44. Xiong, A.S., R.H. Peng, J. Zhuang, F. Gao, Y. Li, Z.M. Cheng, and Q.H. Yao, *Chemical gene synthesis: strategies, softwares, error corrections, and applications.* Fems Microbiology Reviews, 2008. **32**(3): p. 522-540.
45. Withers-Martinez, C., E.P. Carpenter, F. Hackett, B. Ely, M. Sajid, M. Grainger, and M.J. Blackman, *PCR-based gene synthesis as an efficient approach for expression of the A+T-rich malaria genome.* Protein Engineering, 1999. **12**(12): p. 1113-1120.
46. Wu, G., J.B. Wolf, A.F. Ibrahim, S. Vadasz, M. Gunasinghe, and S.J. Freeland, *Simplified gene synthesis: A one-step approach to PCR-based gene construction.* Journal of Biotechnology, 2006. **124**(3): p. 496-503.
47. Xiong, A.S., R.H. Peng, J. Zhuang, J.G. Liu, F. Gao, J.M. Chen, Z.M. Cheng, and Q.H. Yao, *Non-polymerase-cycling-assembly-based chemical gene synthesis: Strategies, methods, and progress.* Biotechnology Advances, 2008. **26**(2): p. 121-134.
48. Dong, B., R. Mao, B. Li, Q. Liu, P. Xu, and G. Li, *An improved method of gene synthesis based on DNA works software and overlap extension PCR.* Molecular Biotechnology, 2007. **37**(3): p. 195-200.
49. Li, S.L. and M.F. Wilkinson, *Site-directed mutagenesis: A two-step method using PCR and DpnI.* Biotechniques, 1997. **23**(4): p. 588-90.
50. Rei, A.P., *Integrating gene optimisation in Plasmodium LysRSs genes for optimal expression in Escherichia coli*, in *Biology Department*. 2010, University of Aveiro. p. 32.
51. Santos, M.A.S., V.M. Perreau, and M.F. Tuite, *Transfer RNA structural change is a key element in the reassignment of the CUG codon in Candida albicans.* Embo Journal, 1996. **15**(18): p. 5060-5068.

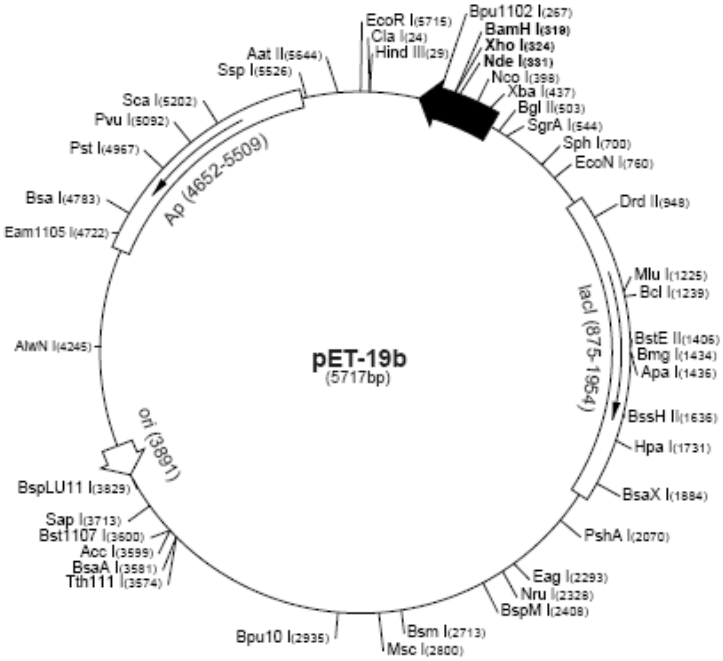
52. Chaubey, S., A. Kumar, D. Singh, and S. Habib, *The apicoplast of Plasmodium falciparum is translationally active.* Molecular Microbiology, 2005. **56**(1): p. 81-89.
53. Woese, C.R., G.J. Olsen, M. Ibba, and D. Soll, *Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process.* Microbiology and Molecular Biology Reviews, 2000. **64**(1): p. 202-+.
54. Behura, S.K., M. Stanke, C.A. Desjardins, J.H. Werren, and D.W. Severson, *Comparative analysis of nuclear tRNA genes of Nasonia vitripennis and other arthropods, and relationships to codon usage bias.* Insect Molecular Biology, 2010. **19**: p. 49-58.
55. Frugier, M., T. Bour, M. Ayach, M.A.S. Santos, J. Rudinger-Thirion, A. Theobald-Dietrich, and E. Pizzi, *Low Complexity Regions behave as tRNA sponges to help co-translational folding of plasmodial proteins.* Febs Letters, 2010. **584**(2): p. 448-454.
56. Herr, A.J., N.M. Wills, C.C. Nelson, R.F. Gesteland, and J.F. Atkins, *Drop-off during ribosome hopping.* Journal of Molecular Biology, 2001. **311**(3): p. 445-52.
57. Angov, E., P.M. Legler, and R.M. Mease, *Adjustment of codon usage frequencies by codon harmonization improves protein expression and folding.* Methods Mol Biol, 2011. **705**: p. 1-13.
58. Matthews, B.W., *The structure of E. coli beta-galactosidase.* Comptes Rendus Biologies, 2005. **328**(6): p. 549-556.
59. Moura, G.R., M. Pinheiro, A. Freitas, J.L. Oliveira, J.C. Frommlet, L. Carreto, A.R. Soares, A.R. Bezerra, and M.A.S. Santos, *Species-specific codon context rules unveil non-neutrality effects of synonymous mutations*, in *unpublished data*. 2011: University of Aveiro.

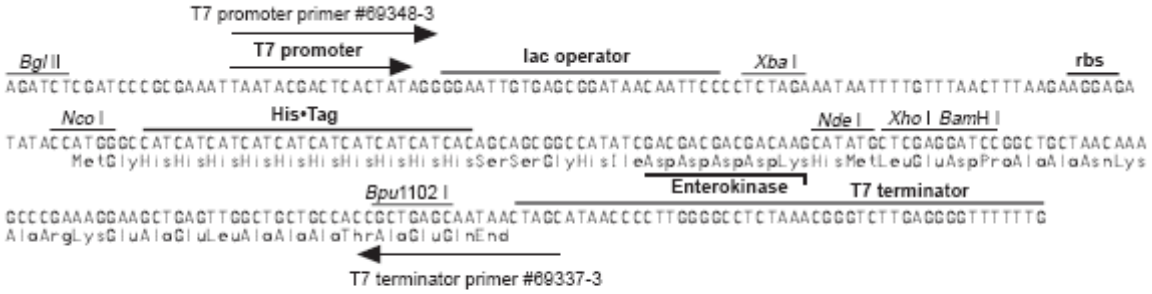
Appendices

APPENDIX A

Plasmids

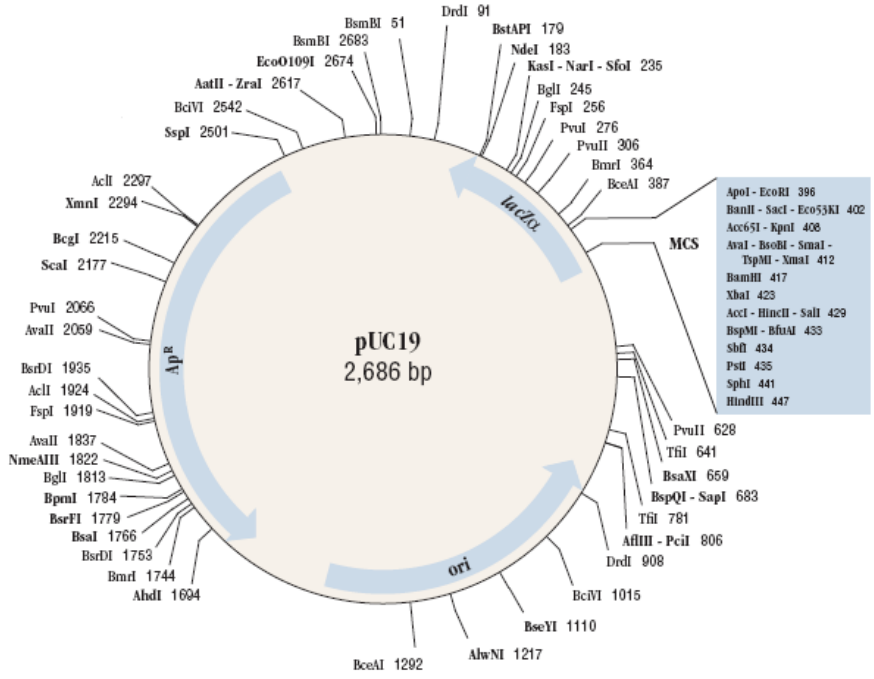
- ***Plasmid pET19b:*** ~57kb*; Ori* – origin of replication for *E. coli; multiple* cloning sites (NdeI-BamH I); T7 promoter; His-Tag coding sequence (10x His); *lac*I coding sequence; enterokinase site.



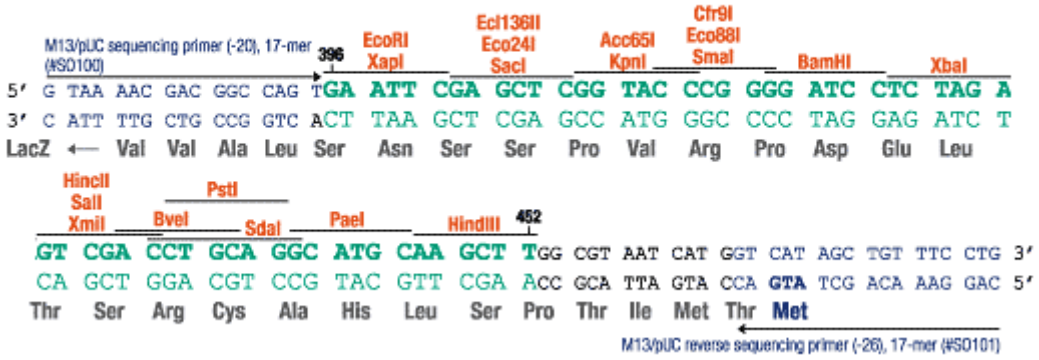


pET-19b cloning/expression region

- ***Plasmid pUC19:*** ~2686 kb. *E. coli* plasmid cloning vector containing portions of pBR322 and M13mp19; high-copy number; Ap^R (Ampicillin resistance gene); *lacZ*α (N-terminal fragment of β-galactosidase).

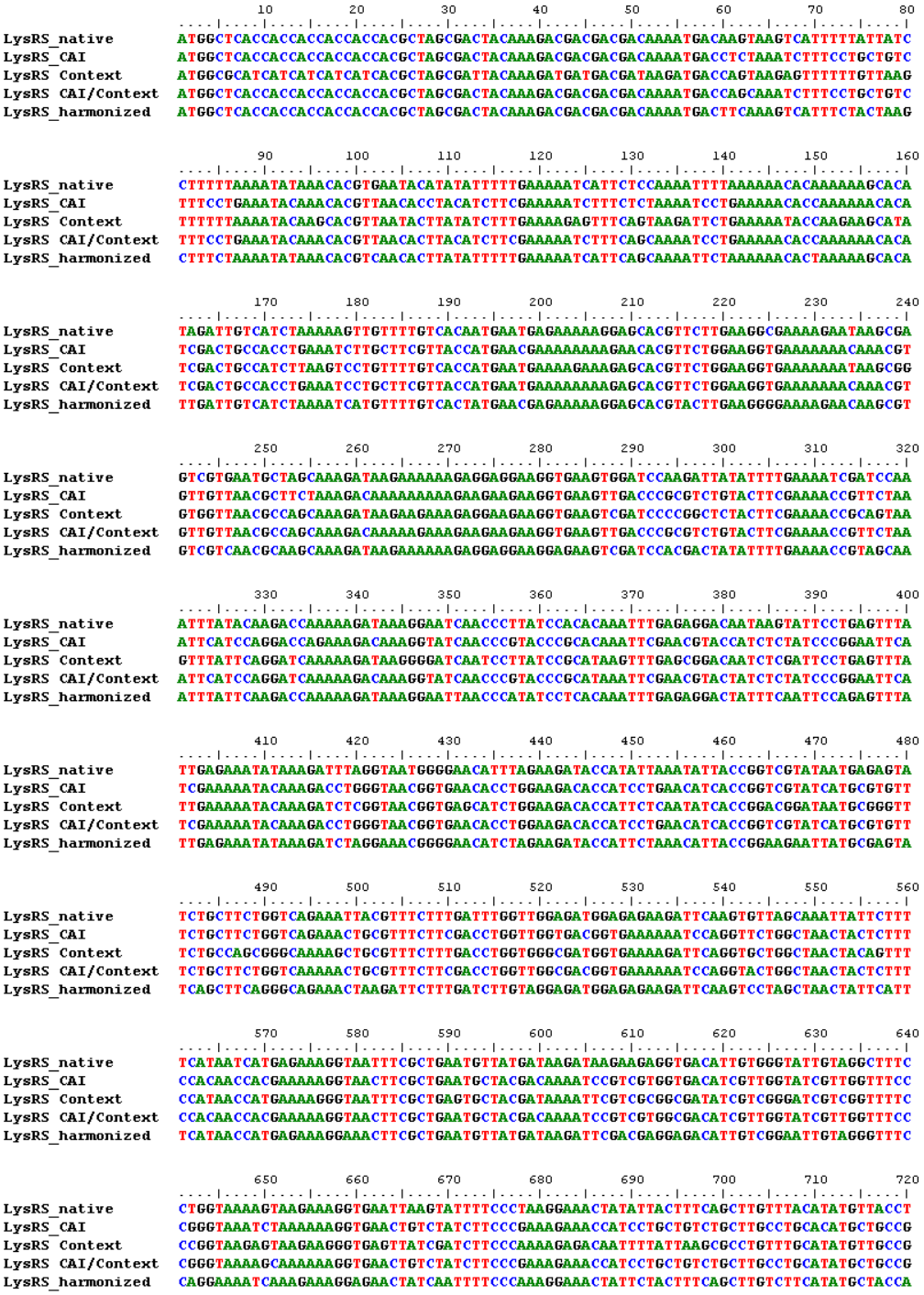


Multiple cloning sites of pUC19

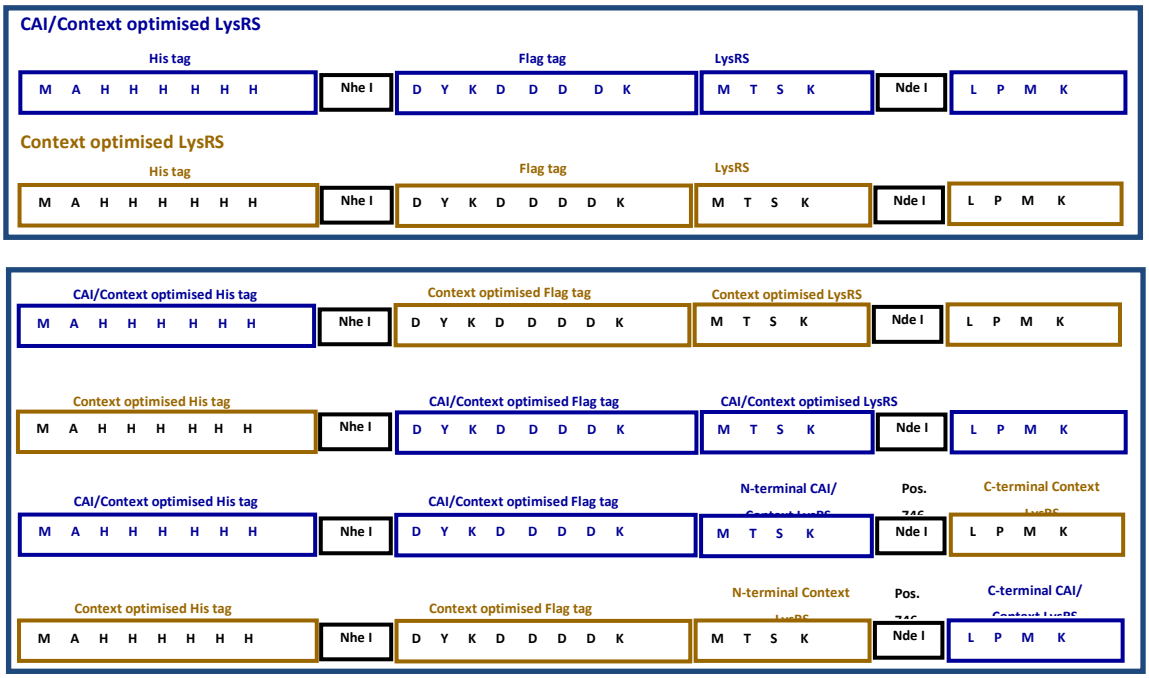


APPENDIX B

- ***Scheme representing the Pf LysRS native sequence and the respective CAI, context, CAI/context, harmonised -optimisations*** *(only 720bp of 1820 are shown).*

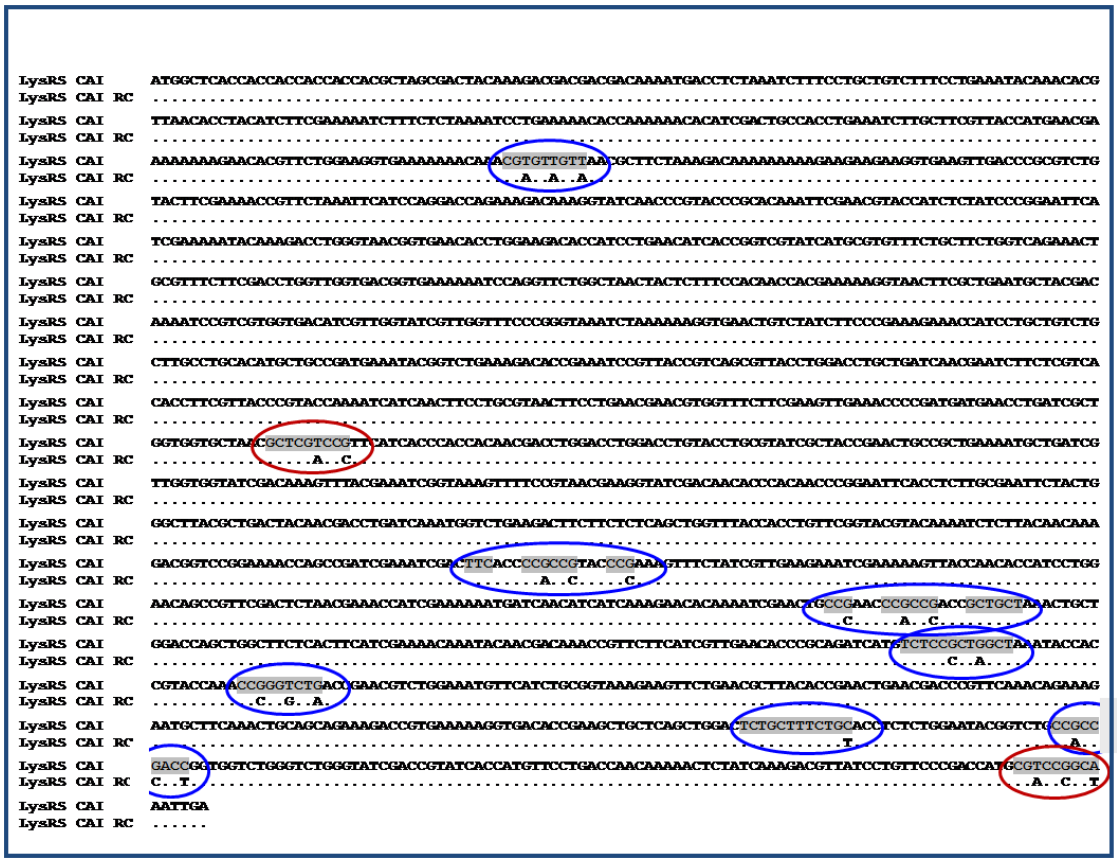


- ***(previous work [50])*** *Hybrid genes identities between: CAI and Context (CAI/CON) - optimised LysRS gene and Context-optimised LysRS gene, with NdeI and NheI restriction sites.* ***Digest 2 (DIG2):*** *Hybrid gene identity used in this study. Note: The DIG2 is equal to the context-optimised gene but the his-tag was changed to a CAI/CON-optimised one.*

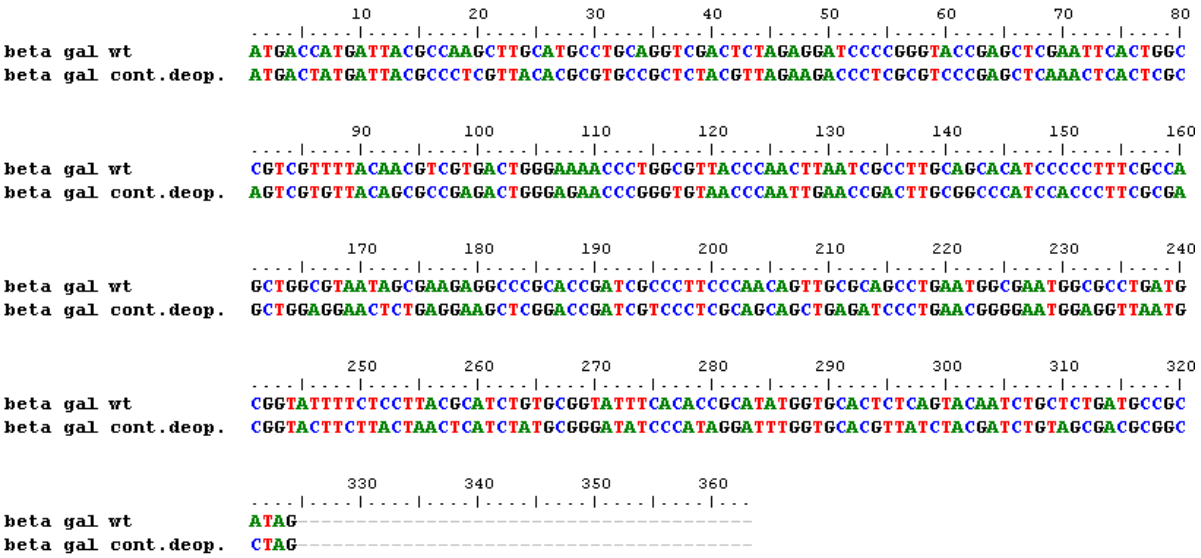


(DIG 2→)

- ***Site-directed mutagenesis of the CAI-optimised LysRS:*** *The nine sites that were targeted by SDM for the introduction of RCRRs are encircled. Red circles highlight sites that could not be mutated. Blue circles indicate mutations that were introduced successfully.*



- ***LacZα gene context-de-optimisation:*** *Alignment between the native and Context-de-optimised gene.*



- ***Schemes showing the differences between the bad context of native and de-optimised gene, obtained in ANACONDA®.*** *Informations represented by row:*
    - ***CAI*** *– codon adaptation index;*
    - ***CU*** *– Codon usage;*
    - ***Nucleotide sequence of the gene*** *- red shadow represents the de-optimised context ("bad" context);*
    - ***AA*** *– Aminoacid sequence.*

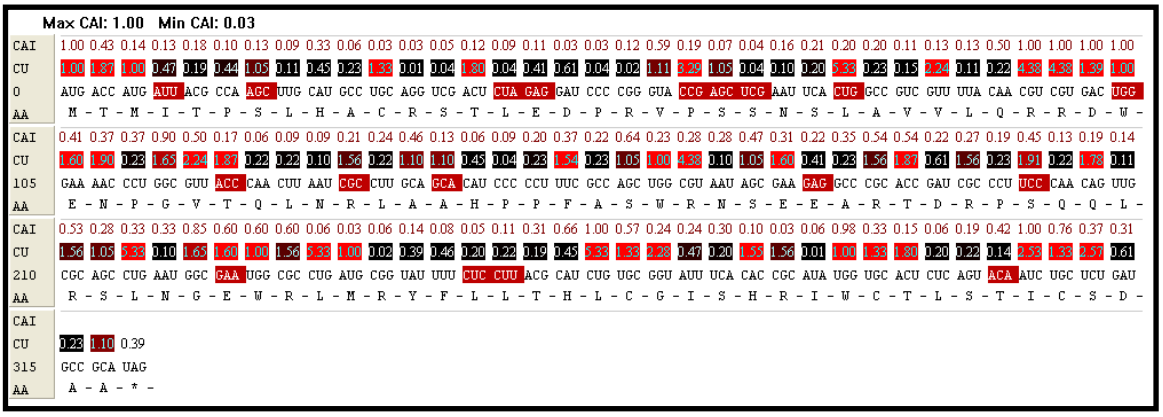


Fig. 19 - LacZα native gene.

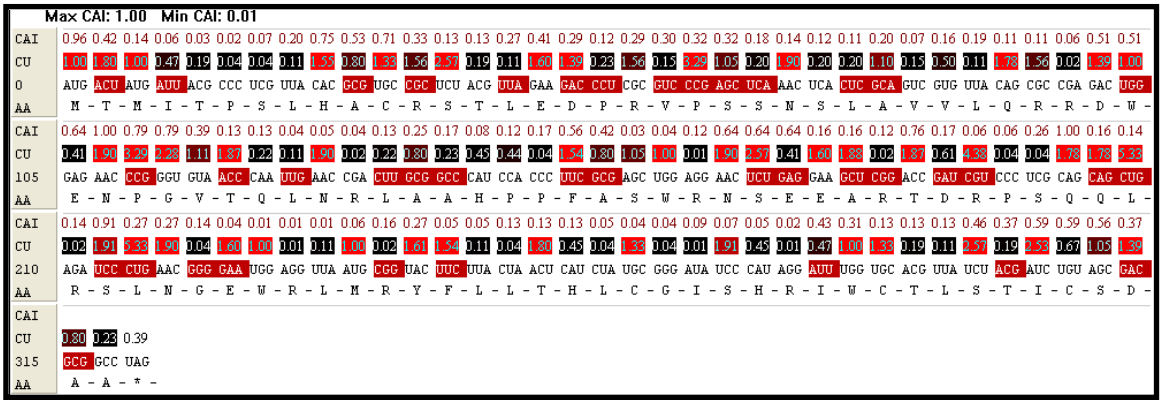
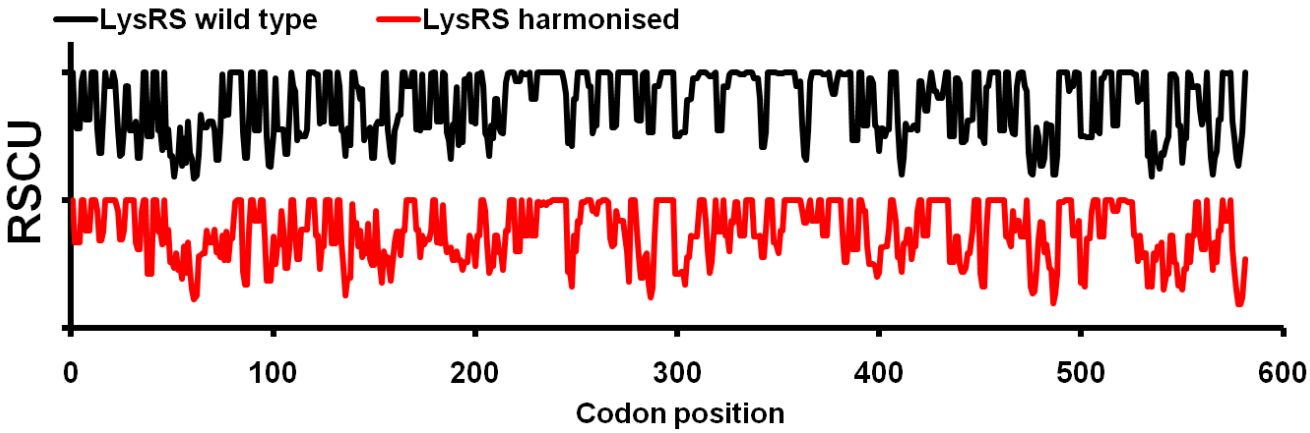


Fig. 20 - LacZα de-optimised gene.

APPENDIX C

- ***Relative Synonymous Codon Usage*** *(RSCU) of the native Pf LysRS gene, comparing with the harmonise construct, designed in this project.*



APPENDIX D

E. coli β-galactosidase three-dimensional structure [58]:

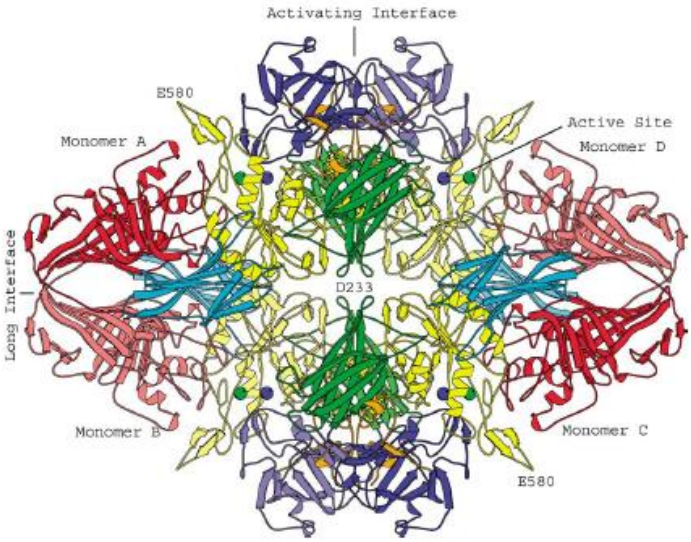


Fig. 21 – The β-galactosidase tetramer. Coloring by domain: complementation α-peptide, orange; Domain 1, blue; Domain 2, green; Domain 3, yellow; Domain 4, cyan; Domain 5, red.

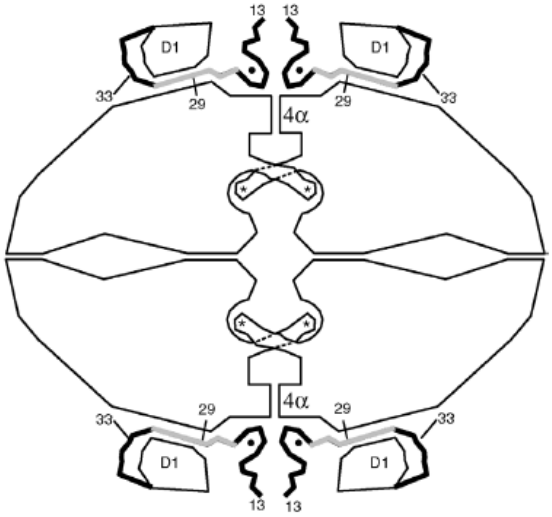


Fig. 22 – Scheme summarizing key features of β-galactosidase. A magnesium ion (shown as a small solid circle) bridges the complementation peptide to the rest of the protein. The four active sites are labeled with asterisks.

APPENDIX E

Procedures:

- ***LB agar plates and liquid medium***

1. LB medium was dissolved at a concentration of 25g/L and 2% Agar in Milli-Q water. (Milli-Q water refers to ultrapure laboratory water that has been filtered and purified by reverse osmosis).
2. The flask for sterilization of the medium was autoclaved;
3. Once the medium reached room temperature ampicillin was added from a stock solution 100mg/ml to a final concentration of 75µg/ml.

Note: The procedure for LB liquid medium is equal without adding Agar

- ***DNA concentration***

1. With the sampling arm open, 1µL of elution buffer from the GeneJET^{TM} or Plasmid Miniprep Kit or QIAquick PCR Purification Kit from Qiagen was pipetted as a blank onto the lower measurement pedestal.
2. Sampling arm was closed and a spectral measurement at 260nm was initiate.
3. When the measurement was completed, the sampling arm was openned and the sample from both upper and lower pedestals was wipe.

The same steps with each sample were repeated and the λ value and the 260/280 ratio for purity analysis were noted.

- ***Digestion with NcoI and XhoI***

Table 1. Volumes for double digestions with NcoI and XhoI (example).

Components	pET 19b
DNA (µl)	10 (89,7 ng/ µl)
10X Buffer (µl)	4
Enzyme (µl) NcoI + XhoI	1
MQ (µl)	4

- ***Digestions with NedI and XhoI***

Table 2. Volumes for digestion with NedI (example).

Components	pUC19
DNA (µl)	19,94 (100,3 ng/ µl)
10X Buffer (µl)	4
Enzyme (µl) XhoI	2
MQ (µl)	14,06

Table 3. Volumes for digestion with NedI (example).

Components	pUC19
DNA (µl)	46 (100,3 ng/ µl)
10X Buffer (µl)	6
Enzyme (µl) NedI	3
MQ (µl)	5

- ***Treatment with shrimp alkaline phosphatase***

1. 1 U of SAP is needed for every 1 µg of plasmid

2. Incubation at 37 ºC for 30-60 minutes.

3. Reaction was stopped by heating at 65 ºC for 15 minutes.

4. To remove any excess of non reacted enzyme, purification using QIAquick PCR Purification Kit Protocol from Qiagen was performed followed by DNA concentration determination in the NanoDrop^{TM} 1000

5. The reactions were then incubated overnight for 16 ºC.

- ***Site Directed Mutagenesis***

DNA template: [CAI optimised-LysRS in pET19b] = 150,5 ng/µl

Working solution required : 2,5 ng/µl

Table 4. Setup volumes for site directed mutagenesis PCR

Component	1	2	3	4	5	6	7	8	9
10xBuffer (µl)	2,5	2,5	2,5	2,5	2,5	2,5	2,5	2,5	2,5
dNTPs (µl)	0,5	0,5	0,5	0,5	0,5	0,5	0,5	0,5	0,5
Primer 1 (µl)	0,64	1,18	0,82	0,95	1	1,03	1,08	1,08	1,05
Primer 2 (µl)	0,64	1,18	0,82	0,95	1	1,03	1,08	1,08	1,05
DNA template(µl)	1	1	1	1	1	1	1	1	1
MQ (µl)	78,88	74,56	77,4	76,4	76	75,84	75,36	75,36	75,6

All the components were mixed with respective volumes and 2 µl of enzyme were added. PCR was performed for 18 cycles.

- ***Plasmid purification (GeneJET™ Plasmid Miniprep Kit)***

For **low-copy** plasmids: 10 ml of culture.

1. The pelleted cells were resuspended completely (by vortexing or pipetting up and down until no cell clumps remain) in 250 µl of the Resuspension Solution and the cell suspension transferred to a microcentrifuge tube.
2. 250 µl of the Lysis Solution were added and mixed thoroughly by inverting the tube 4-6 times until the solution becomes viscous and slightly clear.
3. 350 µl of the Neutralization Solution were added and and mixed immediately and thoroughly by inverting the tube 4-6 times.
4. A centrifugation step was performed for 5 min to pellet cell debris and chromosomal DNA.
5. The supernatant was transferred to the supplied GeneJET™ spin column by decanting or pipetting, without disturbing or transferring the white precipitate.
6. After a centrifugation step for 1 min, the flow-through was discarded and the column placed back into the same collection tube.
7. 500 µl of the Wash Solution were added and the tubes centrifuged for 30-60 seconds. After that the flow-through was discarded and the column placed back into the same collection tube.
8. The wash procedure was repeated (step 7) using 500 µl of the Wash Solution.
9. After discarding again the flow-through, an additional centrifugation step was done for 1 min to remove residual Wash Solution.
10. GeneJET™ spin column was transferred into a fresh 1.5 ml microcentrifuge tube.
11. 50 µl of the Elution Buffer were added to the center of GeneJET™ spin column membrane to elute the plasmid DNA, incubated for 2 min at room temperature and centrifuged for 2 min.
12. The Spin column was then discarded and the purified plasmid DNA store at -20°C.

- ***E. coli competent cells preparation (TFB method)***

Solutions:

TDBI (100ml)

- 0,3g Potassium acetate
- 1,2g $RbCl_2$
- 0,147g $CaCl_2$ or 0,195g $CaCl_2.2H_20$
- 1,0g $MnCl_2$

The listed reagents were dissolved in $d2H_20$ and 17,7ml Glycerol 87% was added. pH was adjusted to 5,8 with 0,2M acetic acid.
Solution was filtered and stored in 2,5ml aliquots at -20ºC.

TFBII (100ml)

- 0,24g MOPS Na
- 1,1g $CaCl_2$ or 1,457g $CaCl_2,2H_20$
- 0,12 $RbCl_2$

Reagents were dissolved in $d2H_20$ and 17,7ml Glycerol 87% added. pH was adjusted to 6,5 with 0,5M acetic acid NaOH. Solution was filtered and stored in 2,5ml aliquots at -20ºC.

Protocol:

1. 200µl cells were inoculated from a overnight culture (5ml), in 5ml LB medium and incubated at 37ºC with 180 rpm until obtain OD_{550}=0,3 (a 2h).
2. Approximately 4ml of anterior culture were inoculated in 100ml LB and grown at 37ºC with 180rpm until obtain OD_{550}=0,3 (a 3h).
3. Cells were collected for two 50ml falcons and putted 5min on ice.
4. The falcon tubes were centrifuges at 2500rpm for 5min at 4ºC.
5. The supernatant was decanted and 2 pellets resuspended in 20ml TFBI (cold) each.
6. Centrifugation was performed at 2500rpm for 5min at 4ºC.
7. The supernatant was again decanted and 2 pellets resuspended in 5ml TFBII (cold) each.
8. Incubation was performed on ice for 5min.
9. 200µl aliquots were distributed for each cold Eppendorf and frozen at -80ºC

APPENDIX F

Solutions recipes:

- ***SOC*** *(100 mL)*

2 g Tryptone;
0,5 g yeast extract;
0,05 g NaCl;
1 mL KCl (250 mM)

pH was adjusted to 7,0 with NaOH (0,5 M) and distilled water added up to 80 mL. After autoclave, 20 mL of sterile filtered (0,22 μm) 1 M glucose were added.

- ***Resuspension Buffer*** *(100mL)*

0,5 M NaCl (2,922 g);
20 mM Tris-HCl (0,242 g);
5 mM Imidazole (0,034 g);
6 M Urea (36,036 g)

pH was adjusted to 7,9 with HCl before sterile filtration.

- ***PBS 1x*** *(1000 mL)*

8 g NaCl;
0,2 g KCl;
1,44 g Na_2HPO_4;
0,24 g KH_2PO_4;
800 mL of distilled water

pH was adjusted to 7,4 with HCl and sterilized by autoclaving.

- ***SDS Running buffer 1X***

25 mM Tris, 192 mM glycine, 0.1% SDS for 2L of 1X Running Buffer

- 28,8 g glycine
- 6,04 g Tris base
- 2 g SDS
- 1,8 L dH_20

Tris base and glycine were dissolved in 1.8 L of dH_2O; SDS was added and mixed; dH_2O was added to a final volume of 2 L.

- ***TGM 1x***

25mM Tris-Base (3,03 g/L)
193 mM Glycine (14,4g/L)
20% Methanol

- ***TBS 1x*** *(1000ml)*

6,05 g Tris;
8,76 g NaCl;
800 ml of distilled water.
The pH was adjusted to 7,5 with 1 M HCl and the volume adjusted to 1 L with distilled water.

- ***TBS-T*** *(1000 ml)*

0,5 ml of Tween-20 in 1L of TBS buffer.

- ***Trypsin (*** *v511, Promega):*

1 ampoule resuspended in 2 ml of Buf1 with MQ;

- ***Coomassie Brilliant Blue (R-250)***
  - 40% Methanol
  - 10% Glacial Acetic Acid
  - 0.05% Coomassie Brilliant Blue R-250