# Modeling Ribosome Dynamics to Optimize Heterologous Protein Production in *Escherichia coli*

S. K. Vu, A. A. Belloti, C. J. Gabriel, H. N. Brochu, E. S. Miller, D. L. Bitzer, M. A. Vouk
North Carolina State University
Raleigh, NC, USA

**Abstract - Ineffective heterologous protein synthesis has often been ascribed to codon bias and rare codons. New experimental evidence suggests that codon bias alone may not be the sole cause of poor translation. In this paper we present a free-energy based model of translation elongation to predict and optimize genes for expression in *E. coli*. The model takes into account second order free energy effects from the binding between the anti-Shine-Dalgarno sequence of the 3' terminal 16S rRNA tail and the mRNA, tRNA abundance, and ribosome displacement. The model and software allow optimization of genes for increased (or decreased) protein yield. The model's predictive and optimization accuracy was assessed by optimizing and expressing three model genes and multiple mRNA variants coding for GST (26 kDa Glutathion S-Transferase from *Schistosoma japonicum*). Protein yield of optimized genes showed increase from their wildtype levels. Optimization of Glutathion S-Transferase from *Schistosoma japonicum* and Alcohol Dehydrogenase from *Clostridium ljungdahlii* DSM 13528 are discussed as examples. Corresponding author, S. K. Vu, can be reached at skvu@ncsu.edu.**

*Keywords* - **Gene Optimization, Heterologous Protein Expression, Codon Bias, Free Energy Signal, Ribosome Displacement**

## I. Introduction

Translation is a biological process by which an organism produces specific polypeptides that fold into functional proteins [1,13]. Heterologous protein synthesis uses translation to produce proteins not normally produced in the host organism [2,3]. Unfortunately, attempts to translate unmodified exogenous genes in production organisms, such as *Escherichia coli*, often result in low or no synthesis of the desired protein [1,2]. Some of the issues identified are poor translation, non-optimal ribosome binding site (RBS), RBS and start codon spacing, frameshifting, and premature termination [1,2]. Although maximizing protein yield has been studied in detail for some time, the underlying processes and effects of translation elongation on protein yield have not been resolved. For many years, the primary cause of low protein yield was thought to be simply codon bias and rare codons coding for rare aminoacyl-tRNA (aa-tRNA) [1,2,3]. Recent experimental evidence suggests that low protein yield may result from a number of equally important additional factors [4,5,6,7].

## II. Protein yield

The current "standard" for determining protein yield is Sharpe's codon adaptation index (CAI) [8]. Sharpe's algorithm takes the codon usage from highly expressed genes as the standard for calculating CAI. Genes with similar codon usage as highly expressed genes score higher, with CAI ranging from 0 to 1. CAI can be used as a measure of a gene's codon bias relative to an organism's codon bias. Typically one would optimize yield by modifying the codon bias of a hetereologous gene towards the codon bias of highly expressed genes of the production organism. Unfortunately, optimization using only CAI sometimes works, and sometimes does not [1,3].

It is suggested from the experiments of [4,5,6,7] that the determinants of low protein yield, or protein yield in general, may not only be from codon bias and rare tRNA availability, but from possibly equally important additional factors. For example, ribosomal profiling data from [6] suggest that the anti-Shine-Dalgarno (aSD) of the 3' terminal nucleotides of the 16S rRNA interacts with the mRNA during translation elongation to "pause" the ribosome. Li *et al* [6] observed little correlation between codon usage/tRNA abundance and ribosome translation speed. They propose that the yield from translation elongation is highly correlated with ribosome pausing at Shine-Dalgarno (SD) like sequences in the mRNA, and that the aSD of the 3' end 16S rRNA tail affects the "speed" of ribosomes translating the mRNA. We conjecture that the ribosome is physically slightly displaced relative to the zero reading frame where ribosome pausing is observed. By being displaced, it takes the ribosome longer to acquire the next aa-tRNA. We have used second order effects from the free energy periodicity signal [9,10,11] and tRNA abundance information [12] to build a novel translation model for *E. coli* that incorporates this fractional displacement notion. Based on our *in vivo* experiments, this model appears to have considerable predictive and optimizing power.

## III. Translation Model And Free Energy Signal

### A. Traditional Translation Model

The biochemistry and mechanisms of translation in microbes has been well studied and is simplified into three stages: initiation, elongation, and termination. At initiation the ribosome subunit forms a complex to begin elongation. During elongation, the ribosome translocates three bases 5' to 3' to decode the mRNA. At termination the ribosome recognizes a release factor at the stop codon and dissociates. Because of space limitations, we refer the reader to [13] for the biochemistry of translation.

### B. Free Energy Signal

It has been recognized for some time that in bacterial translation the 3' terminal nucleotides of the 16S rRNA ("exposed tail" of about 13 nucleotides) continuously interact with the mRNA sequence [6,9,10,11,14,15]. These interactions involve Watson-Crick base pairing where the free energy [16] of the hybridization between the 16S rRNA tail and mRNA can be calculated. **Figure 1** illustrates the average signal that is obtained by calculating hybridization energy at every nucleotide of 200 non-frame-shifting *E. coli* endogenous genes [9]. The most prominent binding energy is at the initiation site, but after that there is a periodic (sinusoidal-like) binding signal (negative free energy indicates binding) that corresponds to "in-frame" ribosome translocation [9,10,11]. When heterologous genes translate without being adapted to the host, this signal is disrupted, which we predict is due to binding misalignments and it becomes more likely

that the yield suffers. The mathematics of the model were previously described in [9,10,11,17,18].

This signal inspired a novel mechanistic model of the process that helps to elucidate frameshifts and to predict and optimize protein yield using second-order energetic interactions between the 16S rRNA 3' terminal "exposed tail" and the mRNA. The model's predictive and optimization power was compared to that of simply using codon bias manipulation and was tested *in vivo*.

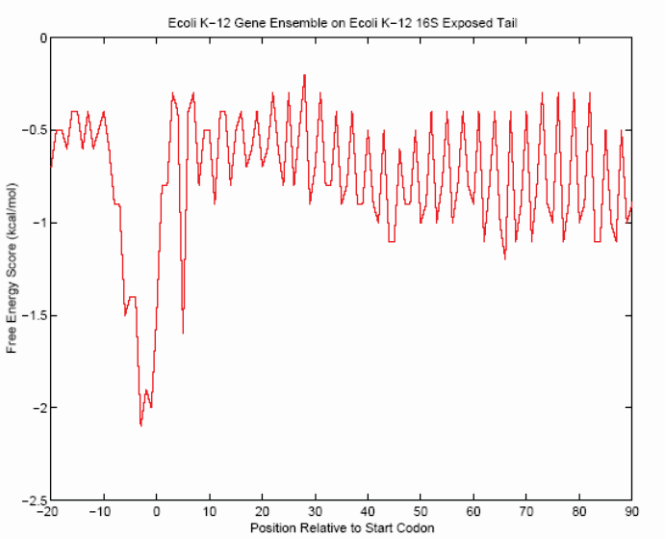


**Figure 1.** Periodic signal observed from average free energy of 200 *E. coli* genes [9].

### C. *The "Spring" Model*

As the ribosome translocates along the mRNA the 16S rRNA 3' terminal end interacts with bases upstream of the ribosome [6,15] (**Figure 2**). If that binding energy is sufficiently large and is on the "wrong" side of favoring normal in-frame binding, it may exert extra force on the ribosome and displace it. In the extreme case, a frameshift may occur. We model this "spring-like" reaction force using a sinusoid. For example, in the +1 programmed frameshifting gene *prfB* (**Figure 2**), the aSD of the 16S tail binds to the SD-like sequences 3 bases (UAU in the figure) upstream from the P site [15] resulting in -9.5 kcal/mol of binding energy. This binding is too close to the P site and compresses the distance between the tail and the ribosome. It behaves like a compressed spring, which upon relaxation displaces the ribosome to minimal energy ¾ of a base downstream into the +1 reading frame. The displaced ribosome then picks up the aa-tRNA in the +1 frame. This leads to a one base frameshift after which the new "in-frame" is maintained [9,10,18]. In contrast, the SD in the -1 frameshifting gene of *dnaX* is 10 bases away from the P site [14]. This binding extends the "spring" and displaces the ribosome upstream to produce a "partial" frameshift (between reading frames; producing both τ and γ subunits). We postulate that a "relaxed spring" state occurs when the "optimal" spacing between the SD and start codon is 8 bases [20]. **Figure 3** shows the *E. coli* 16S aSD during in-frame translation of *lacZ*. The binding energies around that site range from zero to -1 kcal/mol. Because the aSD is bound 8 nucleotides away from the P site, the "spring" is in its "relaxed" state. While slight misalignments between the zero reading frame and A site are possible (and this can affect the yield), there is not enough "spring force" to cause a frameshift.

We model the compression or extension of the "spring" as a change in the "phase" of the free energy sinusoidal signal [10,18] (**Figure 2** and **Figure 3**). In the case of a compressed spring (*prfB*), the phase angle changes 230 degrees at the codon 26 stop [9,18]. In our model, we defined the perfectly relaxed "spring" to have a phase angle of -25 degrees. This is the average phase angle of all non-hypothetical, non-putative, and non-pseudo "long" endogenous genes (1000+ bp) in *E. coli*. We chose to use the phase angle of long genes because we believe that long genes require almost perfect alignment to translate full-length mRNA without errors, i.e. errors can accumulate for very long genes. Therefore the "spring" needs to be close to "relaxed" throughout translation elongation. We define this phase angle as the "species angle".

Ribosome displacement is cumulative and does not reset after translocation. The more the ribosome is displaced, the longer it takes to choose between the two aa-tRNAs of the two reading frames. tRNA abundance of codons in the two reading frames has a major impact on the ribosome wait time. The tRNA abundance translates into tRNA arrival time in our model. When explicit tRNA abundance is not known, we use the codon distribution of the host organism [12]. After the ribosome has "chosen" the next aa-tRNA, it translocates three bases downstream. In our model, one displacement unit is a misalignment of half a nucleotide. Two displacement units is a misalignment of a full nucleotide or a reading frameshift. Effective analysis tools to determine yield are polar, ribosome displacement, and ribosome wait time plots.

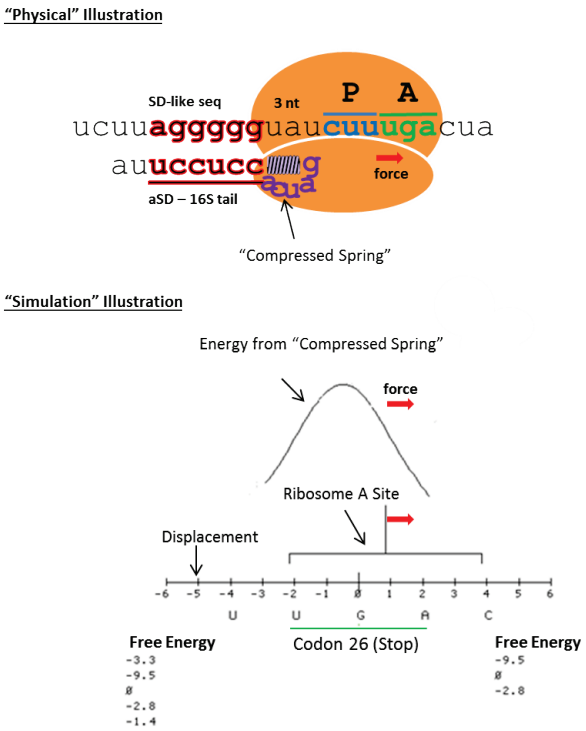


**Figure 2.** Translation simulation of *prfB*. The aSD of 16S "exposed tail" binds too close, 3 nucleotides, to the P site at codon 26 (stop) and compresses the "spring" that displaces ribosome towards minimal energy.

## IV. Results

### A. *Predicting Yield*

Expected yield is proportional to the rate at which ribosomes initate translation and the rate that ribosomes finish translating the mRNA [1]. Our model predicts the ribosome wait time at each codon by calculating the number of cycles it takes to load the next aa-tRNA.

The number of "cycles" at each codon is computed as a function of force from the "spring," ribosome displacement magnitude, and tRNA abundance. We use the the total cycle count as an internal index for measuring protein yield. However, since the overall goal is to minimize displacement throughout the coding region, we will use total displacement change between wildtype and optimized gene to illustrate optimization extent. Total displacement (TD) is the sum of the displacement in absolute value at each codon throughout the coding region.

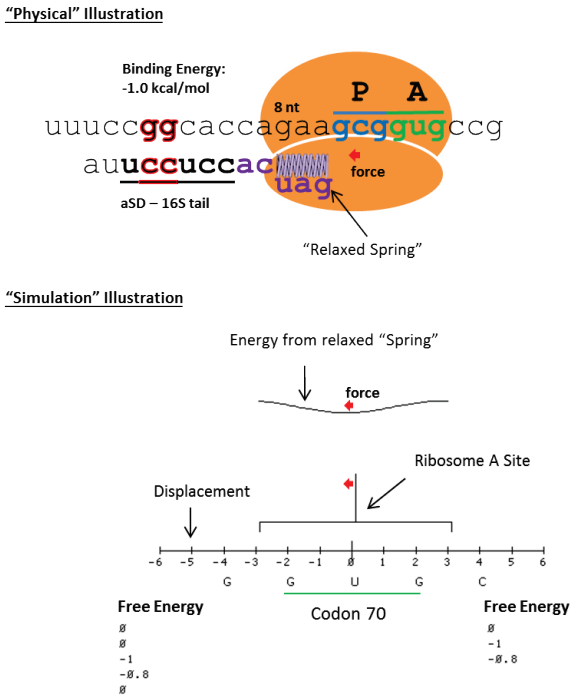


**Figure 3.** Translation simulation of *lacZ* at codon 70. The aSD binds 8 nucleotides from the P site resulting in a "relaxed spring" close to minimal energy with little to no ribosome displacement.

## B. *Optimizing Protein Yield*

We postulate that if we keep the displacement of a gene close to zero and maximize tRNA arrival time by using the most abundant tRNA, we can decrease ribosome wait time and increase protein yield. A displacement close to zero ensures that the A site is aligned close to the zero reading frame. This can be accomplished by choosing the codon that codes for the most abundant tRNA while keeping the phase angle of the gene close to the "species angle". This results in minimal "spring" compression or extension during translocation. Gene designs, incorporating these concepts, are made by changing the genetic sequence using synonymous codons while conserving the amino acid sequence. We believe the model excels at optimizing very long genes, because it minimizes error accumulated by optimizing on "species angle" rather than codon bias. Conversely, genes can be altered to make protein production "less optimal". A decrease in protein yield, with potential applications in pathway optimization and production of "toxic" protein, can be accomplished by keeping displacement further away from 0 but between -1 and +1 to avoid possible frameshifts. This increases the ribosome wait time at each codon.

## C. *In vivo Experiments*

To verify the model's prediction and optimization, we optimized and expressed in *E. coli* three model genes and multiple mRNA variants coding for GST (26 kDa Glutathion S-Transferase from *Schistosoma japonicum*, found in pET-41a(+) plasmid; Novagen, Inc.) Protein yield of optimized genes showed increase from wildtype levels. Due to space limitations, we only illustrate optimization results of *gst* and *adh* (alcohol dehydrogenase, CLJU_C11880, from Clostridium *ljungdahlii* DSM 13528). The first 90 bases of all variants were the same as the first 90 bases of the wildtype. This was done to eliminate variations in yield due to translation initiation [1,5,7]. Genes were cloned into pBAD inducible plasmids (Invitrogen, Inc.) and expressed in *E. coli*. Total protein was quantified by BCA assay and used for normalization of activity; total protein units were measured in absorbance at 562 nm. CAI calculation and optimization were done using the published method [19].

*gst* was optimized based on the displacement (model-optimized), and codon bias (CAI-optimized); see **Figure 4**. All *gst* variants were cloned into the pBAD inducible plasmid, expressed in *E. coli* at 0.02% w/v arabinose for 2 hours, and activity levels assayed. GST activity was measured using E.C. 2.5.1.18 assay. Normalized GST yield was quantified in units of Δ340nm/min per total protein. Three independent inductions were conducted to test for replicability. From each induction three samples were collected for a total of nine samples assayed.

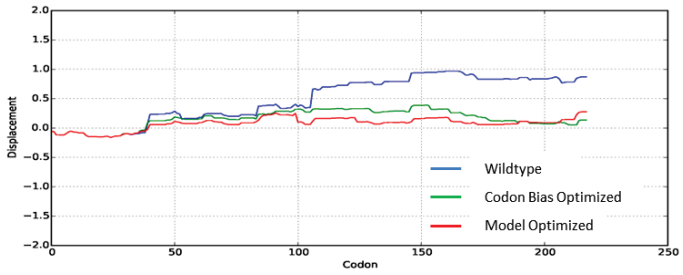


**Figure 4**. Ribosome displacement plots of wildtype (top curve), codon bias optimized (middle curve), and model-optimized (bottom curve). Both optimizations reduce displacement.

The model-optimized variant showed the most GST activity yield followed by CAI-optimized and then wildtype (**Figure 5**). CAI-optimized, which has a better codon bias, did not produce as much GST as model-optimized variant. Conversely, the model optimization of *gst* barely increased codon bias but surpassed the protein yield level of CAI-optimized. Thus, higher codon bias may not always mean higher yield, and CAI is not always an accurate predictor of protein yield as demonstrated by our results and [4,5].

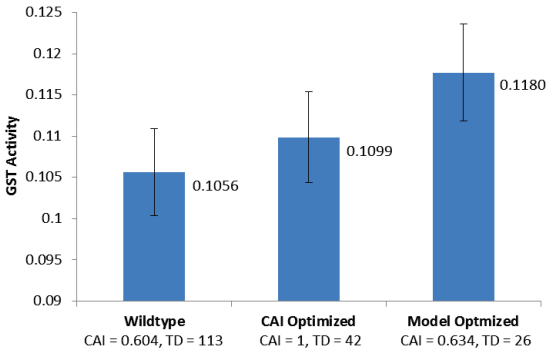


**Figure 5.** GST activity of wildtype and optimized genes. The model-optimized variant yields about 11.6% more GST activity than wild type. CAI-optimized yielded 4% more than wildtype. GST activity units are Δ340nm/min/562nm. Error bar indicates one

standard deviation. CAI is codon adaptation index. TD is total displacement.

*adh* was also optimized based on ribosome displacement (model-optimized; see **Figure 6**). Wildtype and optimized variants were cloned into pBAD inducible plasmid, expressed in *E. coli* at 0.2% w/v arabinose for 4 hours, and activity levels compared. ADH activity was measured using E.C. 1.1.1.1 assay. Normalized ADH yield was quantified in units of Δ340nm/min per unit of total protein. Two independent inductions were conducted to test for replicability. From each induction two samples were collected for a total of four samples assayed. Model-optimized variant yielded a 45% increase from wildtype while barely increasing the codon bias (**Figure 7**).

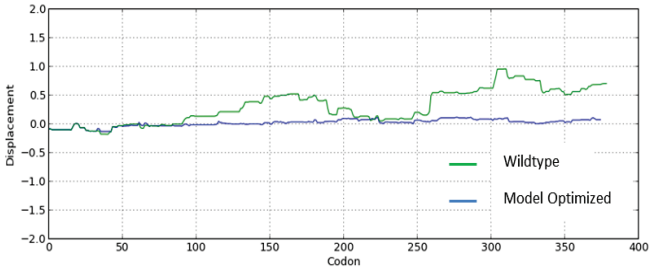


**Figure 6**. Ribosome displacement plots of wildtype (top curve) and model-optimized (bottom curve) *adh*.

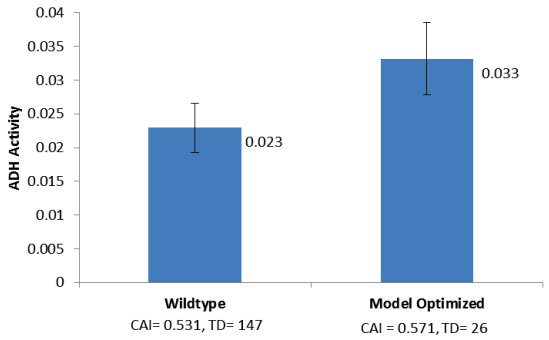


**Figure 7.** ADH activity of wildtype and model-optimized genes. Optimized variant yields 45% more activity than wildtype. ADH activity units are Δ340nm/min/562nm. Error bar indicates one standard deviation. CAI is codon adaptation index. TD is total displacement.

## V. Conclusion

We have developed a new model for predicting and optimizing genes for heterologous protein production. The model incorporates an energetic "spring" of 16S rRNA tail and mRNA interactions, ribosome displacement, and tRNA abundance, leading to a ribosome "wait time" parameter for the gene(s) of interest. This represents a comprehensive strategy for evaluating ribosome dynamics and translational efficiency. The model exists as a fully implemented software package (RiboScan™) that provides a new approach to protein production engineering; this software package will be available as a webservice. Future applications using the model include: 1) analysis of endogenous *E. coli* genes and genome annotations (in preparation), 2) optimization and expression of high value industrial and theurapetic genes that showed poor translation even with codon bias optimization, 3) modifying ribosome "wait time" parameter to optimize protein folding for decreased or elimination of inclusion body formation [6], and 4) expanding the translation model to different production organisms.

## Acknowledgement

This work was funded in part by NC State Distinguished University Research Professorship funds. We would like to thank Tori Jefferson for standardizing GST assays, Eric Whitmire and Tyler Cross for help in coding part of the model, and Liana Lin for gene analysis.

## References

[1] Plotkin, J. B. & Kudla, G. "Synonymous but not the same: the causes and consequences of codon bias." *Nature Reviews Genetics,* 12(1), 32-42 (2010).

[2] Gustafsson, C., Minshull, J., Govindarajan, S., Ness, J., Villalobos, A. & Welch, M. "Engineering genes for predictable protein expression." *Protein expression and purification*, 83(1), 37-46, (2012).

[3] Gustafsson, C., Govindarajan, S. & Minshull, J. "Codon bias and heterologous protein expression." *Trends in biotechnology,* 22(7), 346-353, (2004).

[4] Welch, M., Govindarajan S., Ness J., Villalobos, A., Gurney, A. & Minshul, J. "Design Parameters to Control Synthetic Gene Expression in Escherichia coli." *PLoS on*e, 4(9), e7002, (2009).

[5] Kudla, G., Murray, A. W., Tollervey, D. & Plotkin, J. B. "Coding-sequence determinants of gene expression in Escherichia coli." *Science*, 324(5924), (2009).

[6] Li, G. W., Oh, E. & Weissman, J. S. "The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria." *Nature*, 484(7395), 538-541, (2012).

[7] Allert, M., Cox, J. C. & Hellinga, H. W. "Multifactorial determinants of protein expression in prokaryotic open reading frames." *Journal of molecular biology*, 402(5), 905-918, (2010).

[8] Sharp, P. M. & Li, W-H. "The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications." *Nucleic Acids Res.*, 15(3), 1281-1295, (1987).

[9] Mishra, M., Vu, S. K., Bitzer, D. L. & Vouk, M. A. "Free energy periodicity in prokaryotic coding and its role in identification of +1 ribosomal frameshifting in the Escherichia Coli K-12 gene prfb." 26th *Conf. Proc. IEEE EMBS*, Vol 2, pp. 2848-2851, (2004).

[10] Mishra, M. "The Role of Free Energy Synchronous Signal in Translation of Prokaryotes." Thesis. (www.lib.ncsu.edu/resolver/1840.16/1221) (2004).

[11] Rosnick, D., Bitzer, D., Vouk, M. & May, E. "Free energy periodicity in E. coli coding." *22nd Conf. Proc IEEE EMBS*, Vol. 4, pp. 2470–2473, (2000).

[12] Dong, H., Nilsson, L. & Kurland, C. G. "Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates." *Journal of molecular biology*, 260(5), 649-663, (1996).

[13] Aitken, C. E., Petrov, A. & Puglisi, J. D. "Single ribosome dynamics and the mechanism of translation." *Annu Rev Biophys.*, 39:491-513, (2010).

[14] Larsen, B., Wills, N. M., Gesteland, R. F., Atkins, J. F. "rRNA-mRNA base pairing stimulates a programmed-1 ribosomal frameshift." *J. Bacteriol.*, 176(22), 6842-6851, (1994).

[15] Weiss, R. B., Dunn, D. M., Dahlberg, A. E., Atkins, J. F. & Gesteland, R. F. "Reading frame switch caused by base-pair formation between the 3'end of 16S rRNA and the mRNA during elongation of protein synthesis in Escherichia coli." *EMBO J.*, 7(5), 1503 (1988).

[16] Freier, S. M., Kierzek, R., Jaeger, J. A., Sugimoto, N., Caruthers, M. H., Nielson, T. & D. H. T. II. "Improved free-energy parameters for predictions of RNA duplex stability." *Proc. Nat. Acad. Sci.* USA, 83(24):9373–9377, (1986).

[17] Ponnala, L., Stomp, A.-M., Bitzer D. L. & Vouk M. A. "Analysis of free energy signals arising from nucleotide hybridization between rRNA and mRNA sequences during translation in Eubacteria." *EURASIP J. on Bioinformatics and Systems Biol.*, pp. 1-9 (23613), (2006).

[18] Ponnala, L., Bitzer, D. L., Stomp, A., & Vouk, M. A. "A computational model for reading frame maintenance." *28th Conf. Proc. IEEE EMBS,* pp. 4540-4543, (2006).

[19] Puigbò, P., Guzmán, E., Romeu, A. & Garcia-Vallvé, S. "OPTIMIZER: a web server for optimizing the codon usage of DNA sequences." *Nucleic acids research,* 35(suppl 2), W126-W131, (2007).

[20] Shultzaberger, R. K., Bucheimer, R. E., Rudd, K. E., & Schneider, T. D. "Anatomy of *Escherichia coli* ribosome binding sites." *Journal of molecular biology*, 313(1), 215-228, (2001).