Shine-Dalgarno Anti-Shine-Dalgarno Sequence interactions and their functional role in translational efficiency of Bacteria and Archaea

Akram Abolbaghaei

Supervisor: Dr. Xuhua Xia

Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
University of Ottawa
In partial fulfillment of the requirements for a
Master's degree from the
Ottawa-Carleton Institute of Biology

Thèse soumise à la
Faculté des Etudes Supérieures et Postdoctorales
Université d'Ottawa
En vue de l'obtention de la maîtrise
L'Institut de Biologie d'Ottawa-Carleton

Abstract

Translation is a crucial factor in determining the rate of protein biosynthesis; for this reason, bacterial species typically evolve features to improve translation efficiency. Biosynthesis is a finely tuned cellular process aimed at providing the cell with an appropriate amount of proteins and RNAs to fulfill all of its metabolic functions. A key bacterial feature for faster recognition of the start codon on mRNA is the binding between the anti-Shine-Dalgarno (aSD) sequence on prokaryotic ribosomes at the 3' end of the small subunit (SSU) 16S rRNA and Shine-Dalgarno (SD) sequence, a purine-rich sequence located upstream of the start codon in the mRNA. This binding helps to facilitate positioning of initiation codon at the ribosomal P site. This pairing, as well as factors such as the location of aSD binding relative to the start codon and the sequence of the aSD motif can heavily influence translation efficiency. The objective of this thesis is to understand the SD-aSD interactions and how changes in aSD sequences can affect SD sequences in addition to the underlying impact these changes have on the translational efficiency of prokaryotes.

In chapter two, we hypothesized that differences in the prevalence of SD motifs between *B. subtilis* and *E. coli* arise as a result of changes in the free 3' end of 16S rRNA which may have led *B. subtilis* and *E. coli* to evolve differently. *E. coli* is expected to be more amenable to the acquisition of SD motifs that do not perfectly correspond with its free 3' 16S rRNA end than *B. subtilis*. Further, we proposed that the evolutionary divergence of these upstream sequences may be exacerbated in *B. subtilis* by the absence of a functional S1 protein. Based on the differences between *E. coli* and *B. subtilis*, we were able to identify SD motifs that can only perfectly base pair in one of the two species and are expected to work well in one species, but not the other. Furthermore, we determine the frequency and proportion of these specific SD motifs that are expected to be preferentially present in one of the two species. Our motif detection is in keeping with the expectation that the predicted five categories of SD that are associated with *B. subtilis* and are expected to be less efficient in *E. coli* exhibit greater usage in the former than latter. Similarly, the predicted category of SD motifs associated with the *E. coli* 16S rRNA 3' end is used more frequently in *E. coli*.

Across prokaryote genomes, translation initiation efficiency varies due to codon usage differences whereas among genes, translation initiation varies because different genes vary in SD strength and location. In chapter 3 we hypothesized that there is differential translation initiation between the genes across 16 archaeal and 26 bacterial genomes. We assessed the efficiency of translation initiation by measuring: i) the SD sequence's strength and position and ii) the stability of the secondary structure flanking the start codon, which both affect accessibility of the start codon

Résumé

La traduction est un facteur déterminant pour le taux de biosynthèse des protéines; pour cette raison, les espèces bactériennes développent généralement des caractéristiques afin d'améliorer l'efficacité de leur processus de traduction. La biosynthèse est un processus cellulaire minutieusement régulé afin de fournir à la cellule une quantité appropriée de protéines ainsi que les ARN lui permettant de remplir toutes ses fonctions métaboliques. Une caractéristique bactérienne clé pour une reconnaissance plus rapide du codon de départ sur un ARNm est la liaison entre la séquence anti-Shine-Dalgarno (aSD) sur les ribosomes procaryotes à l'extrémité 3 'de la petite sous-unité (SSU) ARNr 16S et celle de Shine-Dalgarno (SD), une séquence riche en purine située en amont du codon de départ de l'ARNm. Cette liaison contribue à faciliter l'alignement du codon d'initiation avec le site P ribosomique. Cette association, ainsi que des facteurs tels que l'emplacement du site de liaison à une séquence aSD par rapport au codon de départ et la séquence du motif aSD peuvent fortement influencer l'efficacité de la traduction. L'objectif de cette thèse est de comprendre les interactions SD-aSD et comment les différences entre les séquences aSD peuvent affecter des séquences SD en plus de l'impact sous-jacent de ces changements sur l'efficacité de la traduction chez les procaryotes.

Dans le chapitre deux, nous avons émis l'hypothèse que les différences entre *B. subtilis* et *E. coli* dans la prévalence des motifs SD sont les conséquences de changements dans la libre extrémité 3 'de l'ARNr 16S qui pourraient avoir conduit *B. subtilis* et *E. coli* à évoluer différemment. *E. coli* devrait être plus susceptible que *B. subtilis* d'acquérir des motifs SD ne correspondant pas parfaitement à sa libre extrémité 3 'ARNr 16S. En outre, nous avons proposé que la divergence évolutive de ces séquences en amont peut être exacerbée dans *B. subtilis* par l'absence d'une protéine fonctionnelle S1. Sur la base des différences entre *E. coli* et *B. subtilis*, nous avons pu identifier des motifs SD pouvant parfaitement s'apparier dans seulement l'une ou l'autre des deux espèces des motifs SD et qui devraient ne fonctionner correctement que dans une seule des 2 espèces. Par ailleurs, nous avons déterminé la fréquence et la proportion de ces motifs SD spécifiques qui sont censés être présents préférentiellement dans l'une des deux espèces. Notre détection de motif est conforme à la prédiction que les cinq catégories de SD prédites associées à *B. subtilis* et qui devraient être moins efficaces dans *E. coli* présentent une plus grande utilisation

dans la première que dans la dernière. De même, la catégorie prédite de motifs SD associés à l'extrémité 3’ de l'ARNr 16S de *E. coli* 3 'est plus souvent utilisée dans *E. coli*.

Parmi les génomes procaryotes, l'efficacité d'initiation de la traduction varie en raison des différences d'utilisation des codons tandis que, parmi les gènes, l’initiation de la traduction varie parce que des gènes différents n’ont pas la même force ni le même emplacement pour leurs motifs SD. Dans le chapitre 3, nous avons émis l'hypothèse qu'il y a une initiation de la traduction différentielle entre 16 archéobactéries et 26 génomes bactériens. L’initiation de la traduction s’est prouvée plus efficace chez les bactéries Gram-positives que chez les bactéries Gram-négatives mais aussi plus efficace chez les *Euryarchaeota* que chez les *Crenarchaeota*. Nous avons évalué l'efficacité de l'initiation de la traduction en mesurant : i) la force et la position du motif SD et ii) la stabilité de la structure secondaire flanquant le codon de départ, qui affectent tous deux l'accessibilité du codon de départ.

Acknowledgements

My heartfelt thanks go to my supervisor, Dr. Xuhua Xia, an exceptional teacher and model that I have sought to emulate throughout my studies. Under his tutelage and with his unending support, I have been able to enter the realm of research through work in his lab as his master's student. His support has been consistent and constant throughout my graduate studies and I am grateful for his invaluable help with my thesis.

I would like to also express my deepest thanks to Dr. Guy Drouin and Dr. Linda Bonen whom provided significant guidance for the successful completion of this project. Dr. Marie-Andrée Akimenko, Dr. Ashkan Golshani and Dr. Tim.Xing have also been incredibly helpful by accepting to be my thesis examiners: I extend to them also my deep gratitude.

My journey towards the completion of this thesis is also due to productive collaboration and discussion with past and present lab members and peers. Among the latter, Jordan silke and Juan wang were instrumental with their suggestions and direction upon multiple reviews of my thesis. This network has proven invaluable to me and I am proud to be part of it.

I am also indebted to my mother, Nazgol, and my older sister Azam: without these women encouraging me, loving me and holding me up to their expectations and standards, I would not have been able to complete this thesis. My husband Rasool has given me constant strength, prayers and love. His positivity and support account for the basis of much of this thesis's work.

Finally, as a student of the University of Ottawa's Department of Biology, I am thankful for the scholarship and assistantship of this institution of which I am proud to be a member.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

A	Adenosine
C	Cytosine
G	Guanine
T	Thymine
U	Uracil
I	Inosine
Y	Pyrimidines (U/T and C)
R	Purines (A and G)
N	A, C, U/T and G

DNA	Deoxyribonucleic Acid
RNA	Ribonucleic Acid
mRNA	Messenger RNA
rRNA	Ribosomal RNA
tRNA	Transfer RNA
PTC	Peptidyl Transferase Centre
EF-G	Elongation factor G
EF-Tu	Elongation Factor Tu
CDS	Coding Sequences
CF	Codon Frequency
HEG	Highly Expressed Genes
LEG	Lowly Expressed Genes
SD	Shine-Dalgarno
aSD	Anti-Shine-Dalgarno
RBS	Ribosome Binding Site
RPS1	Ribosomal Protein S1

UTR	Untranslated Regions
SSU	Small Subunit
PSD	Proportion of SD-Containing Genes
MSD	Mean Number of Consecutively Matched Sites
TIR	Translation Initiation Region
TTR	Translation Termination Region
NCBI	National Centre for Biotechnology Information
DAMBE	Data Analysis in Molecular Biology and Evolution
ITE	Index for Translation Elongation
MFE	Minimum Folding Energy
DB	Downstream Box
eLF	Eukaryotik Translation Initiation Factors
IF	Translation Initiation Factors
ORF	Open Reading Frame
$tRNA^{fMet}$	Formyl-Methionine tRNA

Amino Acid Abbreviations:

Arg	Arginine
Gly	Glycine
Gln	Glutamine
Leu	Leucine
Met	Methionine
fMet	Formyl-methionine
Pro	Proline
Ser	Serine
Phe	Phenylalanine
Trp	Tryptophan

Chapter 1

Introduction

1.1 Protein Synthesis

Conversion of information from DNA to protein occurs in a step by step process by which a stretch of DNA is first transcribed into mRNA which is subsequently read by specialized cellular machinery to form protein in a process known as translation. Translation is comprised of four phases: initiation, elongation, termination, and recycling.

Ribosomes are the hub for the translation process. Prokaryotes have 70S subunit ribosomes which can be split into 30S small subunits and 50S large subunits. Translation initiation is the series of events through which the 5' untranslated region (UTR) of the messenger RNA (mRNA) attaches onto the ribosome and is correctly positioned for the ribosome to be at the start codon. The ribosome then scans the mRNA from 5' to 3' and reads triplets of nucleotide bases (codons) where each codon corresponds to an amino acid. Transfer RNA (tRNA) acts as carriers of amino acids to the corresponding ribosome site. At any given instance, the ribosome can hold in it a maximum of three tRNAs. The start codon, predominantly AUG, codes for methionine, hence the first tRNA to be recruited is always a formyl-methionine tRNA (tRNAfMet), commencing the elongation phase. Amino acids on adjacent tRNAs form a chain through peptide bonds. As the ribosome moves, the chain lengthens and a polypetide (primary protein) forms. The process continues until the ribosome reaches a stop codon, terminating the protein synthesis. The ribosomal machinery and proteins are recycled to other protein synthesis sites. Translation initiation and elongation are described in detail below.

1.2 Translation Initiation

Of the four steps, translation initiation has incurred significant evolutionary divergence in bacteria, archaea, and eukaryotes. Bacteria have the simplest machinery for translation initiation. Archaea, on the other hand, have fairly complex machinery for initiation, resembling that of eukaryotes (Londei, 2009).

1.2.1 Ribosome Structure and Function

The ribosome being a center for protein formation from genetic information is a highly conserved protein. Both the 30S and 50S subunits have three tRNA binding sites namely: exit site for deacyl-tRNA (E), peptidyl site for peptidyl-tRNA (P) and aminoacyl site for docking aminoacyl-tRNA (A) (Yusupov *et al.*, 2001). The 30S subunit serves two functions. One, it selects the aminoacyl-tRNA that matches correctly to the mRNA codon, and two, the 30S subunit interacts with the 50S subunit to move tRNA and mRNA by precisely one codon (Carter *et al.*, 2000). In the initiation step, it ensures that the formyl-methionine tRNA (tRNAfMET) is correctly positioned in the P-site and binds only to start codon (X. Q. Wu & RajBhandary, 1997). In the elongation phase, it directly controls translational accuracy by checking incoming aminoacyl-tRNA to determine whether it is correct (cognate), a single mismatch (near-cognate) or lacks interaction (non-cognate) with the mRNA codon (Carter *et al.*, 2000; Brodersen *et al.*, 2001; Ogle *et al.*, 2001). In *Escherichia coli*, the 70S ribosome contains 3 rRNA molecules: 16S, 23S, and 5S, as well as 52 proteins: 21 in the small subunit and 31 in the large subunit (Marquez *et al.*, 2011). The small ribosomal subunit (SSU) is composed of the 16S rRNA (1540 nucleotides) and 21 ribosomal proteins. 23S rRNA molecule (2913 nucleotides) has six interwoven domains and 31 ribosomal proteins that make up most of the large ribosomal subunit (LSU). Archaeal ribosome

rRNA generally resemble bacterial rRNA in both size and structure although genomic studies have shown that they have specific affinities with their eukaryal counterparts.

1.2.2 SSU Ribosomal Proteins

Of the fifteen universal ribosomal proteins (Vishwanath *et al.*, 2004) associated with SSU, five (S2, S3, S4, S14, S15) are globular; and a second group of six (S7, S9, S11, S12, S13, S19) have a globular portion plus long unstructured polypeptide segments extending from the globular core (Brodersen *et al.*, 2002). These SSU universal proteins are involved in: SSU RNA folding, stabilizing folded SSU RNA, constraining or stabilizing tRNAs, structural interactions with other ribosome proteins during translation, and controlling SSU binding to LSU (Brodersen *et al.*, 2002). SSU proteins in the second group have interactions with the mRNA decoding site. The former five actively interact with tRNA binding site and/or mRNA. Proteins S9, S12 and S13 interact with A or P sites of tRNA (Brodersen *et al.*, 2002).

Ribosomal protein S1 in *E. coli* plays a key role of unfolding and docking mRNA onto 30S subunit (Duval *et al.*, 2013). It is therefore a key factor for translation initiation in organisms present with it. In *E. coli*, it has been shown to aid in translation of mRNA with weak SD sequences. However, its presence becomes irrelevant when the strength of the SD sequence is increased (Duval *et al.*, 2013).

1.2.3 Translation Initiation Region (TIR)

Translation is a universal process in prokaryotes and initiation step is the rate-limiting step of translation. The 30S ribosomal subunit recognizes mRNA in two pathways. The first is the 30S subunit complexed with IF1 and IF3 binds to mRNA followed by IF2 and GTP-dependent binding of fMet-tRNAfMet. The other pathway is IF2: GTP: fMet-tRNAfMet complex bound to 30S subunit and recognising mRNA (X. Q. Wu *et al.*, 1996). TIR comprises translation initiation codon

(often AUG but also at times GUG (though rarely UUG, AUU, CUG) (O'Donnell *et al.*, 2001)). Shine-Dalgarno sequence located upstream of initiation codon (Shine *et al.*, 1974) and translational enhancers (TE) which (enhance translation in present/absence of SD sequence) including most abundant class contain A/U-rich sequences at upstream of SD sequence and downstream of initiation codon (Qing *et al.*, 2003; Vimberg *et al.*, 2007).

1.2.3.1 The Role of the Shine-Dalgarno Sequence

Shine-Dalgarno sequence was discovered by Australian scientists John Shine and Lynn Dalgarno in 1974. Shine and Dalgarno observed that the *E. coli* ribosomal binding site contained a substantial part of the sequence 5 ' –GGAGGU-3' located 5' to the AUG start codon .The conservation of this motif across different mRNAs suggested an important role for this domain. Through stepwise degradation of the 3' end of 16s rRNA, Shine and Dalgarno further revealed that the 3' end of 16s rRNA is complementary to the SD. Consequently, they proposed that formation of the mRNA-30S ribosomal subunit complex is formed due to the interaction between the 3' end of 16S rRNA and the domain preceding the start codon in mRNA (Shine and Dalgarno, 1974).

It has been long presumed that translation initiation in prokaryotes dominantly followed the Shine-Dalgarno (SD) dependence mechanism (Myasnikov *et al.*, 2009; Malys *et al.*, 2011). However, recent studies have shown a high proportion of leaderless genes, having short or lacking 5'-UTR in Archaea (Nakagawa *et al.*, 2010), and smaller gene proportions in Bacteria using SD-independent mechanisms (Zheng *et al.*, 2011). The generic SD motif is AGGAGG, although it may vary in length from about 4 -7 bp and is usually found 5-13 bases upstream of translation initiation site, within the ribosome binding site (RBS) (Shine & Dalgarno, 1974; Zheng *et al.*, 2011).

The SD sequence base pairs in a Watson-Crick fashion (Shine & Dalgarno, 1974; Lee *et al.*, 1996), with the complementary anti-SD sequence (aSD) at the 3'end of the 16S RNA, which directly positions the start codon to the P-site of the 30S subunit (Myasnikov *et al.*, 2009) and stimulates translation initiation. Prior to and after initiation, the SD-duplex causes strong anchoring of 5' end of mRNA onto the 30S subunit. After initiation of translation, in presence of SD duplex, the mRNA moves in the 3'-5' direction while simultaneously rotating clockwise and lengthening the SD duplex (Yusupova *et al.*, 2006). In *E. coli*, this clockwise rotation and lengthening brings the SD duplex into contact with ribosomal protein S2 (Yusupova *et al.*, 2006).

1.2.3.2 Role of mRNA Secondary Structure

Minimum free energy (MFE) measures the RNA-formed secondary structure stability expressed in KJ/mol. It is the amount of energy required to break the secondary structure. Stability in the secondary structure increases as the MFE becomes more negative. RNA secondary structure is determined by base pairs that require energy to separate them. To be able to measure the secondary structure stability, different base pairs are assigned different energy indices associated to the strength of the base pair bonds. For example, C/G, A/U and G/U pairs could be assigned values -4, -3, and -2, respectively. Folding energy (FE) is defined as a function of these allowed base pairs, with the the summation of these index values for all base pairs. Therefore, more negative FE indicates more stable secondary structure. MFE is considered as the minimum FE corresponding to the most stable secondary structure. (de Smit *et al.*, 1990).

In order to unwind such strong secondary structure, ribosomes consume a lot of energy and time resulting in a waste of the cell's energy. Several researchers have stated that protein production can decrease due to stable secondary structures in 5' UTR (Osterman *et al.*, 2013). Secondary structure stability can be measured by the strength of base pair bonds among different

nucleotides by examining the change in Gibb's free energy necessary to generate the most stable structure.

We used MFE as a proxy for translation initiation. Tools available for measuring secondary structure stability are Vienna RNA Package (Hofacker, 2003), mfold (Zuker, 2003) and UNAfold software (Markham *et al.*, 2005). We used DAMBE to compute MFE in our study which implements the functionality of Vienna RNA package (Hofacker, 2003). The settings used in Dambe for our analysis are: folding temperature as 37°C, with no lonely pairs and with no G/U pairs at the end of helices. However, any changes in these settings does not affect the relative magnitude of MFE.

1.2.3.3 Nature, Length and Distance to Start Codon of SD Sequence

Since base-pairing of SD-aSD may limit the movement mRNA along the ribosome, the nature, the length of the SD motif, and distance of SD to start codon are key factors in translation initiation (Ringquist *et al.*, 1992; Starmer *et al.*, 2006; Osterman *et al.*, 2013). The nature and length determine the stability and hybridization of the SD-aSD interaction. The formation of hydrogen bonds between aligned, SD and anti-SD nucleotides base pairing in a Watson-Crick fashion form a more stable double stranded structure having a lower free energy than the singular counterparts. In the dynamic cellular environment, SD can base pair either fully or partially to its complement, resulting in variations of SD motifs. The length of the SD motif also affects rate of translation. Varying the length of the *E. coli* SD sequence from five to eight bases results in a fourfold increase in gene expression level (Ringquist *et al.*, 1992). However, a longer SD motif increases SD-aSD interactions and inhibits translation from progressing into elongation phase (Komarova *et al.*, 2002).

The need for a specific distance between start codon and SD sequence for optimal SD-aSD interaction has already been established (Ringquist *et al.*, 1992; Chen *et al.*, 1994; Osterman *et al.*, 2013). The space between the SD and start codon acts as a hinge to facilitate start codon base pairing with tRNAfMet in the P-site (Osterman *et al.*, 2013). Typical SD sequences have the center between 7-15nt to the start codon, with the complementary region ranges from 2-8nt (Osterman *et al.*, 2013).

1.2.3.4 Role of the Start Codon

The start codon is the gateway to initiation of protein production. The nature of the start codon greatly affects translation initiation rate (Ma *et al.*, 2002; Osterman *et al.*, 2013). There are three universal start codons: AUG, GUG, and UUG. Species tend to prefer some start codons over others (Ma *et al.*, 2002). Ma and his colleagues in 2002 examined the Correlations between Shine-Dalgarno Sequences and Gene Features Such as predicted expression levels and operon structures in 30 prokaryotic species including gram positive bacteria, gram negative bacteria, and archaea species. The highest preference among different prokaryotic genome was observed for AUG followed by GUG and finally UUG (Ma *et al.*, 2002). He also observed that genes with AUG start codon tend to have a higher SD% than genes with GUG and UUG start codon. He then suggested that the weak start codon conjunction with lack of SD sequence might reduce the gene expression. Protein yield also follows the same order (Osterman *et al.*, 2013). Studies on SD sequences either through sequence similarity or free energy calculations have shown that genes with AUG are more likely than genes with other start codons, such as GUG or UUG, to possess an SD sequence (Ma *et al.*, 2002; Osterman *et al.*, 2013).

In order to identify SD sequences, it is not appropriate to define the SD sequence simply as an AGGAGG motif located within a fixed distance range upstream of the initiation codon. The

SD sequence located on the mRNA and the anti-SD sequence on the small subunit of ribosome pair to position the anticodon of the initiation tRNA properly at the start codon. Based on several previous studies, the optimal location of SD is often measured as the distance of SD to AUG start codon (e.g. D1 and D2 in Fig. 1.1) called spacing or from the middle of the SD sequence to AUG start codon (called aligned spacing) (Ringquist *et al.*, 1992; Chen *et al.*, 1994; Osterman *et al.*, 2013). Nevertheless, this approach is possibly incorrect as demonstrated in Fig. 1.1, both SD1 and SD2 position the tRNA anticodon properly at the start codon AUG, but their associated D1 and D2 (distance) are different. A correct distance measure should take into consideration the relative position of both rRNA tail and mRNA. One such distance (D_{toAUG}) is the distance in number of nucleotides from the end of the SSU rRNA to the beginning of the start codon.

While numerous studies on the Shine-Dalgarno sequence and its effects on mRNA secondary structure flanking the start codon have been carried out and the research has been used to develop computational tools for optimizing translation initiation (Na *et al.*, 2010) , there is currently no computational index to measure translation initiation.

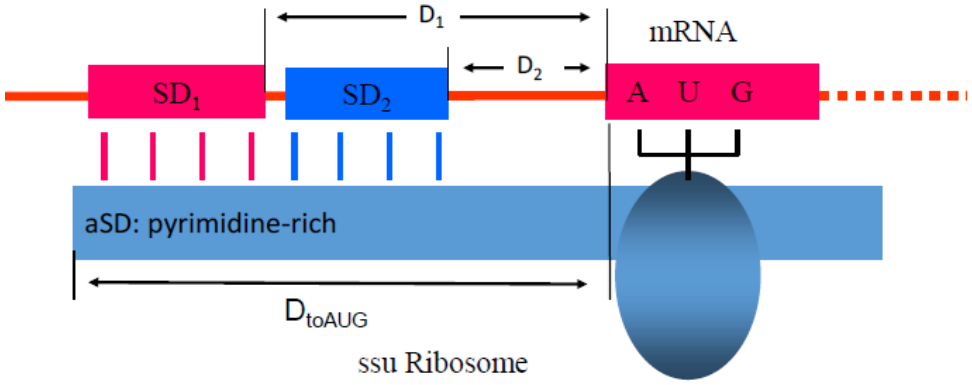

Figure 1.1. Overview of Shine-Dalgarno sequence located on mRNA pairing with anti-SD (aSD) sequence on the small subunit (SSU) rRNA. Figure reproduced from (Prabhakaran *et al.*, 2015) with permission.

1.2.4 Initiator tRNA

Initiator tRNA brings the methionine (code for start codon) to the initiation complex for the initiation of protein synthesis. They are recognized by few initiation factors and discriminated by translation elongation factors. Unique feature of initiator tRNA include the absence of a Watson-Crick base pair (traditional base pairing: A:T or C:G) between positions 1 and 72 in the acceptor stem and three conserved consecutive G:C base pairs in the anticodon stem (There is a missing base pair between position 1 and 72 in the acceptor and G:C base pairs in the anticodon stem) (Cory *et al.*, 1970; Barraud *et al.*, 2008). Methionine formylation by methionyl-tRNA transformylases to tRNAfMet is another important feature of initiator tRNA (Barraud *et al.*, 2008). In addition, unlike elongator tRNA which enter via the A-site and then translocate to the P-site, initiator tRNA can directly bind to SSU at the P-site. Initiator tRNA interaction with the A-site requires the tRNA to bind with elongation factor (EF)-Tu.GTP dimer. However, this does not take place due to the weak binding ability of initiator tRNA with the EF-Tu.GTP dimer (Hansen *et al.*, 1986).

1.3 Translation Initiation in Eukaryotes

In eukaryotes, translation is initiated by a scanning mechanism in which the small ribosomal subunit (40S) bind the 7-methyl guanosine cap structure at the 5'end of mRNA with several initiation factors (Furuichi *et al.*, 1975) and slid downstream to find the first initiation codon (AUG) which is surrounded by a particular sequence such as the Kozak sequence (Kozak, 1999).

1.4 Translation Initiation in Bacteria

In bacteria, translation involves the formation of three intermediary initiation complexes: (1) the 30S preinitiation complex (pre-30SIC) in which there is no physical information taking place between the mRNA and the fMet-tRNAfMet, (2) a 30S initiation complex (30SIC) obtained through a conformational changes of the pre-30SIC leading to the first codon-anticodon interaction in the 30S peptidyl (P) site, and (3) the 70S initiation complex for the elongation phase (Myasnikov *et al.*, 2009; Simonetti *et al.*, 2009). Essentially, initiation factor 1 (IF1) binds onto the 30S ribosomal subunit which triggers 70S subunit dissociation. IF1 then stimulates IF2 and IF3 activity and stabilizes IF-2 binding on 30S ribosomal subunit while preventing tRNA from binding to the 30S aminoacyl (A) site (Myasnikov *et al.*, 2009).

IF2 then interacts with the tRNA which allows fMet-tRNAfMet binding to the 30S. This mediated subunit joining uses GTPase activity. In the final step, IF3 correctly positions the mRNA

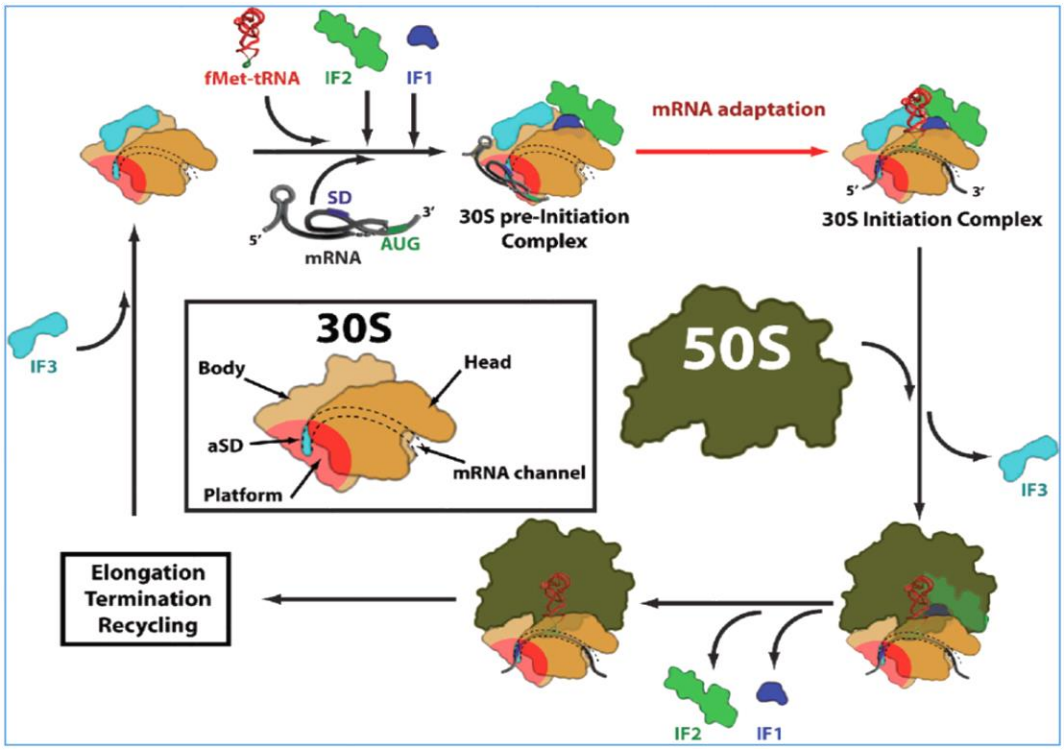

Figure 1.2. An overview of translation process in *Escherichia coli* (Simonetti *et al.*, 2009) with permission.

On to the 30S subunit and encourages codon-anticodon interactions at the ribosomal 30S peptidyl (P) site (Laursen *et al.*, 2005; Myasnikov *et al.*, 2009; Simonetti *et al.*, 2009). This interaction is further stabilized with the help of the Shine-Dalgarno (SD) sequence present in the 5' UTR of the gene, which base pairs in a Watson-Crick fashion (Shine & Dalgarno, 1974; Lee *et al.*, 1996) with the complementary anti-SD sequence at the 3'end of the 16S RNA. In this manner, the start codon is directly positioned at the P-site of the 30S subunit. As illustrated in Figure 1.2, structured mRNA binds to 30S by docking of the mRNA on the platform of the 30S subunit which forms the pre-initiation complex following by the accommodation of the mRNA into the normal path to promote the codon anticodon interaction in the P site [21]. The resulting 30SIC engages

the 50S subunit to form the 70SIC from which the initiation factors are expelled and the synthesis of the encoded protein can proceed through the elongation, termination and ribosome recycling phases (Figure 1.2).

1.5 Translation Initiation in Archaea

Translation initiation in Archaea is a hybrid of features observed in bacteria and eukaryotes. Similarities with bacteria include: (1) the preference of AUG, GUG, and UUG as start codons in the order listed, as well as (2) the use of polycistronic mRNAs (Ma *et al.*, 2002). One major difference, however, is that a large portion of archaeal genes lack altogether a 5'-UTR and the initiation codon is located a few nucleotides after the 5'-end or directly at it (termed leaderless mRNA). Hence there are two mechanisms of translation initiation: SD-dependent for leadered mRNA and SD-independent for leaderless mRNA. In the former, the mRNA undergoes translation similar to that in bacteria while it has been proposed that, in the latter, the translation follows a eukaryotic mechanism (Benelli *et al.*, 2011).

While the nature of archaeal initiation complexes (ICs) is not fully understood, it is thought to follow the eukaryotic manner of translation initiation. The initiator tRNA is carried to the 30S subunit by a/eIF2 and also binds guanosine triphosphate from hydrolysis of tRNAfMet binding to start codon. Two additional factors aIF1 and aIF1A stimulate binding of tRNAfMet and mRNA to SSU. Protein aIF2/5B comes in at a later stage to adjust tRNAfMet in the ribosomal P site and is also thought to promote LSU binding to SSU. Protein aIF6 prevents premature binding of LSU to 30S IC (Benelli & Londei, 2011).

1.6 Translation Elongation

In the elongation phase, an amino-acyl tRNA having an anticodon complementary to the mRNA codon, binds onto the ribosomal A-site. A peptide bond forms between methionine on the

fMet-tRNA located at the P-site and the amino acid on the amino-acyl tRNA on the A site. The fMet-tRNA then leaves the P- site and peptidyl-tRNA at the A-site translocates with the mRNA, from the A-site to the P-site (Grill *et al.*, 2000). Another amino-acyl tRNA having an anticodon complementary to the nucleotide triplets next to the peptidyl-tRNA at the A-site moves onto the A-site. The process of elongation continues in this fashion until the ribosomal subunit reaches a stop codon (UAA, UAG, and UGA) (Grill *et al.*, 2000). Upon reaching the stop codon, termination factors bind to the ribosome and promote hydrolysis of the peptidyl-tRNA. The ribosomes are then recycled through interactions with various protein factors to generate ribosomal subunits that are capable of undergoing another round of translation (Hershey *et al.*, 2012).

1.7 Translation Termination

The termination of protein synthesis is influenced by the type of stop codon and the nucleotides following it. This phase of translation starts when the A-site of the 30S ribosomal subunit encounters any of the stop codons; UAA, UGA and UAG. In *E. coli*, stop codon UAA followed by U have the highest efficiency in translation termination than other combinations (Poole *et al.*, 1995). These stop codons are decoded by protein factors called class 1 release factors (RF). In bacteria, release factor RF1 recognizes the UAA and UAG stop codons and RF2 recognizes UAA stop codons (E. Scolnick *et al.*, 1968; Capecchi *et al.*, 1969; E. M. Scolnick *et al.*, 1969). In *E.coli*, prfA genes encode RF1 and prfB genes encode RF2. Upon recognition of the stop codon, class I release factor stimulates hydrolysis of the ester bond and releases the polypeptide chain from the ribosome complex (Caskey *et al.*, 1968; E. M. Scolnick & Caskey, 1969). Once the peptide chain released, a GTP dependent class-II RF3 simplifies the dissociation of RF1 or RF2 (Freistroffer *et al.*, 1997). Consequently, GTP hydrolysis which results in the

expulsion of RF3 occurs following by a conformational change at the ribosomal complex (Zavialov *et al.*, 2002). This step signifies the end of protein synthesis.

1.8 The Standard Genetic Code

The standard genetic code (SGC) is used by almost all organisms. Standard genetic code transcribes information stored in DNA and translates it into its amino acid. The genetic code decides each codon should be coded by which amino acid. As shown in Figure 1.3, the standard genetic code includes 64 codons. Out of these 64 codons, 61 are recognized as sense codons and encode amino acids. Among sense codons, one of them code for methionine (AUG) which is serves as the translation initiator or start codon. Termination codons are UGA, UAG and UAA. All amino acids can be coded by more than one codon except methionine and tryptophan. The genetic code is known to be 'degenerate' or 'redundant'. Codon families can be categorized into five different types based on the number of codons in each family. Figure 1.3 shows the different codon families called single, two, three, four or six codons. All codons belonging to a codon family vary only in the base at the third codon position with the exception of those encoding Leu and Arg. Since these two codon families are six-fold degenerate, two of the possible six codons vary at the first position in addition to containing either purine or pyrimidine ending codons at their third codon position. The only three codon family is the isoleucine codon family which includes codons ending with all three bases with exception of G at the third codon position. The four codon families are characterized by their conserved first and second positions while their third codon position may be occupied by all four bases. Six codon families are made by combining two and four codon families.

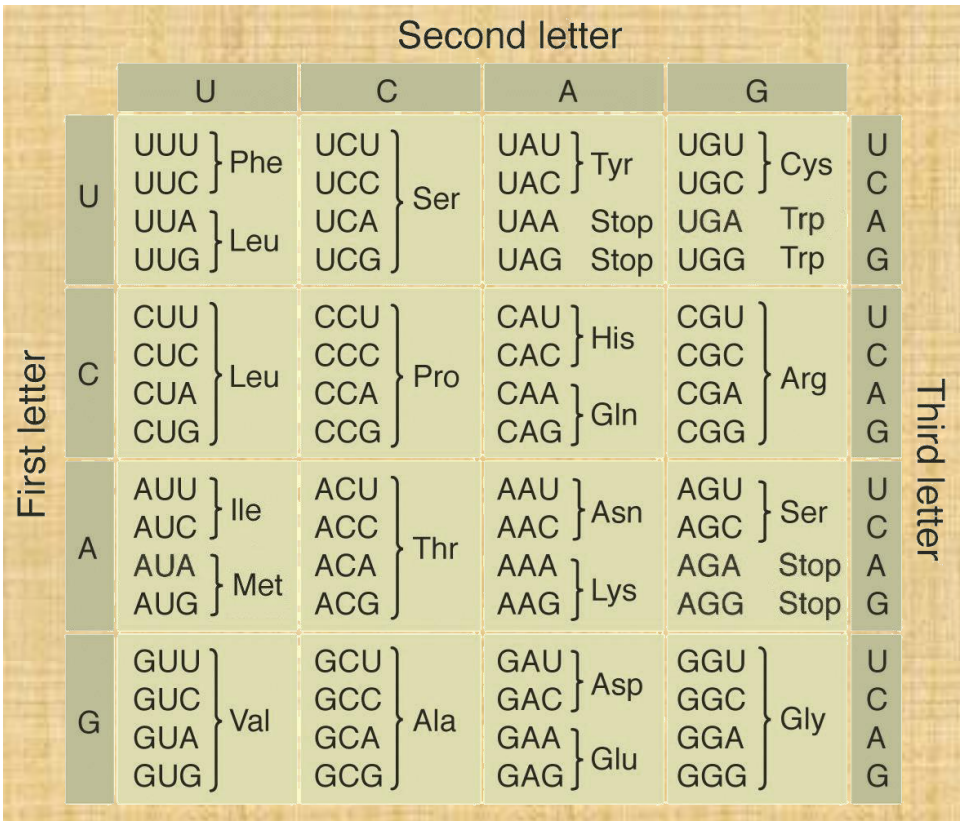

Figure 1.3. Representation of the standard genetic code of 64 codons

1.9 Synonymous Codons

Genetic information on the mRNA template is made up of four nucleotides primarily: guanine, adenine, uracil and cytosine. However, only a combination of three nucleotides code for an amino acid therefore there are 4^3= 64 codons of which 61 code for 20 amino acids. The codons are redundant in that more than one codon can code for the same amino acid (synonymous codons) but are also specific in that none of them encode more than one amino acid (Gouy *et al.*, 1982).

Although there are synonymous codons for an amino acid, some are more abundant than others. There are three hypotheses that have been proposed for explaining why some codons are chosen over other synonymous codons. These hypotheses apply to codon adaptation in: different genomes (Xia, 1996), different genes within the same genome (Gouy & Gautier, 1982), and

different sections of the same gene (Akashi, 1994). These hypothesis include the mutational bias hypothesis (Martin, 1995), the transcription-maximization hypothesis (Xia, 1996) and the translational efficiency hypothesis (Ikemura, 1981b; Bulmer, 1987).

Mutational Bias Hypothesis

When initially proposed, a mutational bias hypothesis was based on observed differences in nucleotide (GC) compositions of mitochondrial genomes which were proposed to be influenced by endogenous DNA damage (Martin, 1995). Generally, low intracellular cytosine (C) levels in bacteria promote pathways that limit C usage (Xia *et al.*, 2006) and a biochemical explanation has been offered which explains the why there is a C-limitation in bacterial species (Rocha *et al.*, 2002). Furthermore, C is susceptible to U/T mutation which also reduces the C use in genome replicates. Naturally, selective pressures would favor against C use and, since every guanine interacts with a cytosine, the consequent use of G is also reduced. Of the three nucleotides in a codon, the third codon is least susceptible to changes and in bacteria, ranges from less than 20% to more than 90% G+C (Muto *et al.*, 1987) and has a higher G+C contents than the prior two nucleotides (Muto & Osawa, 1987). Therefore, G+C variation is a major factor affecting codon variation. Furthermore, in prokaryotes, genomes vary from 25% to 75% G+C content and the length and GC% content of coding sequences (CDSs) are negatively correlated (Xia *et al.*, 2006).

Transcription-maximization hypothesis

The transcription-maximization hypothesis claims that patterns of codon usage that increase transcriptional efficiency increase mRNA concentration which in turn increases the initiation rate and advertently the rate of protein synthesis (Xia, 1996). Here, the differences in codon usage bias among: different genomes, different genes within the genome, and different segments of a gene. These are explained as resulting from selection pressure which favors for

increased translational initiation efficiency which can be increased by mRNA concentration (Xia, 1996). Proof of this proposition is that theoretical as well as empirical data confirm mRNA concentration to be a rate limiting factor in protein synthesis (Xia, 1995).

Translational efficiency hypothesis

Of the three hypotheses, TEH is the commonly referenced hypothesis and has an empirical data supporting it. It underpins strong selection pressure that favors increased protein synthesis and a coding strategy that increases translation efficiency by optimizing initiation and elongation both of which in turn increase protein synthesis rate. Translation is the most energy consuming process and, on the basis of natural selection, we can expect that translation efficiency is subject to high selective pressure. Three ideas support this hypothesis. Firstly, the frequency of codon usage is positively correlated to tRNA availability (Ikemura, 1981b; Gouy & Gautier, 1982). Second, highly expressed genes exhibit greater usage bias than lowly expressed genes (Bennetzen *et al.*, 1982; Sharp *et al.*, 1989). Thirdly, it has been shown that mRNA consisting of preferred codons is translated faster than mRNA selectively modified to contain rare codons (Robinson *et al.*, 1984; Sorensen *et al.*, 1989)**.**

1.10 Translation Efficiency

Efficient translation refers to protein production at a level such that the benefits from the protein exceed the energy cost of production (Dekel *et al.*, 2005). Evolution of cells to tune the efficiency of translation is evident as different genes are expressed in different levels under varying conditions (Meaning that some genes are induced by different environmental factors). Having a standard genome-wide translation scheme thus allows determination of efficiency of translation of various genes expressed in different conditions. Genes, depending on their sequences, are either more or less efficient in utilizing available cellular resources of translation. Translational

efficiency is predominantly determined by both translation initiation and elongation processes (Dekel & Alon, 2005).

1.11 Computation of Translation Efficiency

To understand conditions and pressures that promoted codon-anticodon coevolution and adaptation, a codon-anticodon adaptation (promoted tRNA selection which in this case tRNA pool determine the selection of codon-anti codon) theory has been proposed (Akashi, 1994; Xia, 1998, 2008). It states that: (1) both translation initiation and elongation are rate limiting steps for protein production, (2) tRNA anticodon and codon usage have coevolved simultaneously which has allowed for a correct protein translation (Xia, 1996; Moriyama *et al.*, 1997), and (3) highly expressed genes have a higher codon-anticodon adaptation which promotes elongation efficiency and accuracy (Xia, 1998).

To better understand codon-anticodon adaptation coevolution, several computer models have been generated. However, these either focus on translation initiation (initiation models) or on elongation (elongation models). The initiation models assume that initiation rate alone is the rate limiting step for translation efficiency (Bulmer, 1991; Xia, 1996) while elongation models assume that elongation alone is the rate limiting step for translation efficiency (Bulmer, 1987).

There are several models for calculating initiation or elongation and these include focus on gene-specific codon usage such as: codon adaptation index (CAI) (Sharp *et al.*, 1987; Xia, 2007), translation adaptation index (tAI) (dos Reis *et al.*, 2004) or coding sequences only such as Nc (Wright, 1990) and CDC (Z. Zhang *et al.*, 2012). While the former two have been used extensively for models to measure translation elongation, the latter is not typically related to translation rate.

Codon adaptation index (CAI)

Codon adaptation index (CAI) was first suggested by Sharp and his colleagues to measure the codon usage bias of a gene (Sharp & Li, 1987). Dislike RSCU which is a codon specific index, CAI is a gene-specific index of codon usage bias. The gene whose codon usage bias is to be measured is compared with a reference set (set of genes such as ribosomal RNA that are known to be highly expressed genes of the organism).

Effective number of codons (Nc)

Effective number of codons (Nc) is a codon usage index established by Frank Wright It (Wright, 1990) measures the extent of deviation from even usage of synonymous codons. Nc values can range from a minimum value of 20 considered as extreme bias when only a specific codon in each synonymous codon family has been used to code for an amino acid and up to a maximum value of 61 when there is absolutely no bias when all codons happen at identical frequencies. Nc is considered as a gene-specific index of codon usage bias similar to CAI. The main difference between the two indices is the fact that CAI computation involves a reference set while Nc does not.

Relative synonymous codon usage bias (RSCU)

It is a codon-specific index for codon usage, whereas CAI is a gene-specific index for codon usage. Both are related to gene expression, especially in prokaryotes and unicellular eukaryotes. RSCU-Measures codon usage bias for each codon family. When there is no codon usage bias =1, A codon is overused if its RSCU value is greater than 1 and underused if its RSCU value is less than 1. It is also computed directly from input sequences in Dambe.

Index of Translation Elongation (I_{TE})

Recently, a new translation elongation model has been proposed: the Index of Translation Elongation (I_{TE}), which is similar to the CAI index, except that it takes into account background mutations (Xia, 2015). The I_{TE} computation can be applied in a fashion which treats NNR and NNY subcodon families differently based on the observation that R-ending codons seem to influence codon production more strongly than Y-ending codons (Xia, 2015). This is because genes encoded by the R-ending codons are typically decoded by two types of tRNA species: one with a wobble C and the second with a wobble U while Y-ending codons are decoded typically by a single tRNA species with either a wobble G or A modified to inosine, but never both (Marck *et al.*, 2002; Grosjean *et al.*, 2007).

1.12 Significance of the Study

Genes with higher translation initiation efficiency share two main features: they first possess strong Shine-Dalgarno (SD) sequences on their mRNA which can be found at an optimal position upstream of the start codon, and, secondly, their translation initiation region (TIR) exhibits a weak secondary structure near the positions of the SD sequence and start codon (Osterman *et al.*, 2013). Previously studies have reported observed avoidance of stable secondary structures in start site of mRNA in a widespread range of species. Strong secondary structures are able to mask the translation initiation signals such as SD and start codon from ribosome. Subsequently, it becomes difficult for the ribosome to recognize the Shine dalgarno and start codon (de Smit & van Duin, 1990). Additionally, several previous studies have informed that stable secondary structures in 5' UTR decreased protein production considerably (de Smit & van Duin, 1990; Osterman *et al.*, 2013). They are also able to influence degradation of mRNA. Furthermore, reduced secondary structure patterns were known to be near the translational start site among different cellular species

(Kudla *et al.*, 2009; Tuller *et al.*, 2010). Codon adaptation depends on translation initiation efficiency (Supek *et al.*, 2010; Tuller *et al.*, 2010; Prabhakaran *et al.*, 2015; Xia, 2015). In efficient translation initiation, elongation becomes rate limiting and selection pressure increases translation efficiency and drives codon adaptation. However, when translation initiation is inefficient, selection pressure does not increase translation efficiency because elongation is not rate limiting. This research has also great impact on biotechnological implications such as more efficient insulin production in *E. coli*

Three factors have a strong influence on translation initiation efficiency in bacteria: (1) the nature of the start codon (Hartz *et al.*, 1991; Ringquist *et al.*, 1992; O'Donnell & Janssen, 2001; Ma *et al.*, 2002; Osterman *et al.*, 2013; Prabhakaran *et al.*, 2015), (2) the base pairing strength and position of the Shine–Dalgarno (SD) sequence upstream of the start codon (Shine & Dalgarno, 1974; Hui *et al.*, 1987; de Smit *et al.*, 1994; Olsthoorn *et al.*, 1995; Osterman *et al.*, 2013; Prabhakaran *et al.*, 2015), and (3) the stability of the secondary structure of sequences flanking the SD sequence and start codon (de Smit & van Duin, 1990, 1994; Nivinskas *et al.*, 1999; Milon;Maracci*; et al.*, 2012; Milon & Rodnina, 2012; Osterman *et al.*, 2013; Prabhakaran *et al.*, 2015), with a positive correlation between higher translation initiation and weaker secondary structures of these flanking sequences.

The presence of a SD sequence is positively correlated to gene expression levels (Ma *et al.*, 2002), and it has been proposed that protein production is increased only when translation initiation is efficient. Prokaryotic species, generally require SD binding and a weak secondary structure around start codons for translation initiation. The recognition of a start codon in prokaryotes can occur in four ways: presence of a SD sequence alone, presence of a SD sequence aided by S1, absence of SD sequences but aided by S1 as seen in leaderless open reading frames

(ORFs) in bacteria, and absence of SD and S1 protein as seen in most archaeal genes. Although translation initiation efficiency across prokaryotic genomes varies due to codon usage differences, among prokaryotic genes, it differs based on SD strength and location. These observations prompted us to examine the efficiency of translation among bacteria and archaea and examine their SD-aSD interactions in details and expand our research to 42 species; 26 Bacteria and 16 Archaea. We measured translation initiation by: i) the strength and position of SD sequence and ii) the stability of the secondary structure flanking the start codon, which affects accessibility of the start codon. Our findings imply that there is a differential translation efficiency across the genes in selected species.

Based on our analysis of the differences between 16S rRNA 3' ends in *E. coli* and *B. subtilis*, we looked into the differences between the 16S rRNA 3' ends of *E. coli* and *B. subtilis* and we were also able to identify SD motifs that can only perfectly base pair in one of the two species. Remarkably, our findings highlight the fact that changes in aSD affects SD sequences in both *E. coli* and *B. subtilis*.

Chapter 2

How changes in Anti-Shine-Dalgarno sequence would affect Shine-Dalgarno sequence in *E. coli* and *B. subtilis*

2.1. Abstract

The 3’ ends of the 16S rRNAs in bacteria are directly involved in the selection and binding of mRNA transcripts during initiation of protein synthesis. Translation initiation encompasses well-documented interactions between a Shine-Dalgarno (SD) sequence located upstream of the initiation codon and an anti-Shine-Dalgarno (aSD) sequence at the 3’ end of the 16S rRNA. Consequently, the 3’ end of 16S rRNA (3’ TAIL) is strongly conserved among bacterial species because a change in the region may impact the translation of many protein-coding genes. *Escherichia coli* and *Bacillus subtilis* differ in their 3’ ends of 16S rRNA, being GAUCACCUCCUUUCU in *B. subtilis* and GAUCACCUCCUUA in *E. coli.* Thus, some SDs can base-pair well with aSD in *B. subtilis* (SD_{Bs}) will not base-pair in *E. coli,* and some SDs that some base-pair well in *E. coli* (SD_{Ec}) will not in *B. subtilis.* Selection mediated by the species-specific 3’ TAIL will favour SD_{Bs} against SD_{Ec} in *B. subtilis* but favour SD_{Ec} against SD_{Bs} in *E. coli.* Among well-positioned SDs, SD_{Bs} is used significantly more frequently in *B. subtilis* than in *E. coli,* and SD_{Ec} shows the opposite pattern. Our finding suggests that a change in 3’ TAIL can result in fundamental changes in protein production of many genes.

2.2. Introduction

In the 1980s, the development of a new standard for identifying bacteria was well underway. Woese and his colleagues showed that the best approach to determine phylogenetic relationships between bacteria and other life-forms was to compare segments of the genetic code that have proven to remain relatively stable among lineages (Woese *et al.*, 1985; Woese, 1987; Clarridge, 2004). In bacteria, regions coding for 5S, 16S, and 23S rRNA as well as their intergenic space are the best candidates, as these have the most stable sequences throughout the entire spectrum of the phylogenetic tree (Acinas *et al.*, 2004). A simple reason for this would be that, as these regions are crucial to translation, the integrity of these sequences is critical to cells. Among these three regions, 16S rRNA has been the most used candidate for taxonomic purposes in bacteria (Bottger, 1989; Palys *et al.*, 1997; Kolbert *et al.*, 1999; Tortoli, 2003; Clarridge, 2004) and, not only can it be compared among all prokaryotes but it can also be compared to the 18S rRNA gene of eukaryotes and 16S rRNA gene of archeobacteria (Woese *et al.*, 1985; Woese, 1987; Pace, 1997; Palys *et al.*, 1997; Clarridge, 2004).

The 16S rRNA functions as a facilitating agent in ribosomal interactions with mRNA through improving recognition and binding of mRNA with the initiation complex (Shine & Dalgarno, 1974; Steitz *et al.*, 1975), and is highly reliable for the identification and taxonomic classification of bacterial species through phylogenetic analysis. Because of its ubiquitous nature (Woese, 1987; Orso *et al.*, 1994) (there are often multiple copies located in different operons throughout the genome) and its highly conserved sequence (Woese, 1987; Orso *et al.*, 1994; Clarridge, 2004; Chakravorty *et al.*, 2007), 16S rRNA is considered the most reliable genetic marker to study bacterial phylogeny and taxonomy (Patel, 2001; Janda *et al.*, 2007). As such, random sequence changes associated with the 16S rRNA are being used as a measure of

evolutionary changes (Patel, 2001). Furthermore, the 16S rRNA sequence is usually around 1500 bp, which is sufficient for informatics purposes (Janda & Abbott, 2007).

Although it is a single strand nucleic acid, 16S rRNA molecule has many self-complementary regions which, upon binding, form double-helical regions interspaced with hairpin loops. Its 3' and 5'ends are short single-stranded tails that can pair with other single-stranded nucleic acids. In general, all species adopt the same basic secondary structure of their 16S rRNA which is crucial to perform their function in translation (Woese *et al.*, 1980). Studies suggest that the secondary structure is more highly conserved than the primary structure because, most often, when mutations do occur in the double-stranded region, they are compensatory to maintain this functional structure (Cammarano *et al.*, 1983).

It is the 3' terminal region of the 16S rRNA that initiates the process of translation of an mRNA. Indeed, during the process of translation initiation, selection of the correct start codon and translational reading frame by 16S rRNA on most mRNAs is usually made possible by a pyrimidine-rich sequence in the short single-stranded tail at the 3' terminus of 16S rRNA. What makes the translation initiation possible is the capacity of this tail to base-pair with a purine-rich segment within the initiator region of a natural mRNA (Luhrmann *et al.*, 1981). In genomes, these sequences are usually found within the first 10 nucleotides upstream of a start codon (Steitz & Jakes, 1975; Dunn *et al.*, 1978; T. Taniguchi *et al.*, 1978; Eckhardt *et al.*, 1979; Luhrmann *et al.*, 1981).

On the mRNA, the corresponding purine-rich sequence is located in the 5' untranslated region (UTR) of mRNA and is termed the Shine-Dalgarno (SD) sequence (Shine & Dalgarno, 1974). As the SD sequence pairs with the complementary anti-Shine-Dalgarno (aSD) region on the free 3' end of 16S rRNA, the alignment of the start codon with the decoding P site of the

ribosome is facilitated (Steitz & Jakes, 1975; Calogero *et al.*, 1988; Kaminishi *et al.*, 2007; Korostelev *et al.*, 2007). rRNA components of the ribosome make extensive contact with both the initiator tRNA and mRNA at the P site (Carter *et al.*, 2000; Berk *et al.*, 2006; Korostelev *et al.*, 2006; Selmer *et al.*, 2006). It was suggested that such interactions modulate the fidelity of initiation (Lancaster *et al.*, 2005). In prokaryotes, the correct recognition of the canonical start codon in mRNA is important for establishing the translational reading frame of protein biosynthesis (Hui & de Boer, 1987; Vimberg *et al.*, 2007; Prabhakaran *et al.*, 2015). Because non-canonical (codons other than AUG, GUG and UUG) or inappropriate start codons would result in unwanted and potentially harmful products in the cell, the translation process is tailored to be very discriminatory against these codons (Ringquist *et al.*, 1992; O'Donnell & Janssen, 2001; Osterman *et al.*, 2013). However, the nature and type of start codon in bacterial species can strongly affect translation initiation efficiency (Hartz *et al.*, 1991; Ringquist *et al.*, 1992; O'Donnell & Janssen, 2001; Osterman *et al.*, 2013; Prabhakaran *et al.*, 2015). This is best demonstrated by studies showing that single substitutions affecting position two or three of the canonical start codon AUG (e.g. AUU, AUC and ACG) can reduce the level of detectable protein products by 100-fold *in vivo* (Sacerdot *et al.*, 1996; Sussman *et al.*, 1996; Qin *et al.*, 2007; Qin *et al.*, 2009).

The SD initiation mechanism discovered in *Escherichia coli* has long been regarded as the predominant mechanism for bacterial translation initiation (Shine & Dalgarno, 1974). Translation initiation by the SD sequence has been experimentally demonstrated in both bacteria and archaebacteria (Band *et al.*, 1984). The short SD motif, typically involving GGAGG, can base pair with a complementary aSD sequence (CCUCC) located at the 3'end of 16S rRNA (Shine & Dalgarno, 1974; Gold *et al.*, 1981; Nakagawa *et al.*, 2010). The base pairing of SD with aSD facilitates translation initiation by anchoring the small ribosomal subunit (30S) around the

initiation codon to form the preinitiation complex (Dontsova *et al.*, 1991). It was observed that the number of genes with SD sequence varies greatly between bacteria phyla with, for instance, Firmicutes which possess the largest fraction of genes preceded by SD sequences while Tenericutes, especially mycoplasma species, have a relatively small fraction of genes coupled with SD sequences (Chang *et al.*, 2006). Since 5' UTRs are highly divergent in prokaryotes and heterogeneous with respect to SD content and since even organisms that are deficient in SD are still capable of translation, this demonstrates that more flexible mechanisms of translation initiation are possible.

For instance, a ribosomal protein (Ribosomal Protein S1 or RPS1) has been identified as a crucial element for identification of the translation initiation region on mRNAs in most gram-negative bacteria, especially for genes lacking SD sequences or for those with weak SD sequences (Farwell *et al.*, 1992; Sorensen *et al.*, 1998; Qu *et al.*, 2012; Duval *et al.*, 2013). In particular, S1 can also form a crucial component of mRNA binding sites designed for mRNA lacking or bearing weak SD sequences (Farwell *et al.*, 1992; Komarova *et al.*, 2005; Duval *et al.*, 2013).

On the contrary, gram-positive bacteria have been found to have either no RPS1 or no conserved RPS1. In the latter case, the number of domains found in RPS1 ranges from one to six (*Spiroplasma kunkelii* and *Mycoplasma pulmonis* contain one domain) while gram-negative bacteria have a relatively high abundance of conserved RPS1 formed of six similar domains (Salah *et al.*, 2009). RPS1 has the ability to effectively unwind nucleic acid helices (Kolb *et al.*, 1977). This would allow the 5' UTR of mRNA to become more permissive to the ribosomal interactions which initiate translation regardless of the presence of a SD sequence upstream of a start codon (Tedin *et al.*, 1997). Using Cryo-EM analysis, RPS1 has been found to interact with both 16S rRNA and the region surrounding SD sequences on mRNA (Sengupta *et al.*, 2001).

The base-pairing strength and position of SD sequences (Shine & Dalgarno, 1974; Hui & de Boer, 1987; de Smit & van Duin, 1994; Olsthoorn *et al.*, 1995; Osterman *et al.*, 2013; Prabhakaran *et al.*, 2015), optimal distance to the start codon, and structural accessibility of this aSD-SD interaction are all critically involved in modulating the efficiency of translation initiation and thus gene expression (Barrick *et al.*, 1994; Chen *et al.*, 1994; de Smit & van Duin, 1994; Rinke-Appel *et al.*, 1994; Prabhakaran *et al.*, 2015). More recently, finer details about the importance of translation initiation signals have emerged from studies, emphasizing on the fact that surrounding nucleotides may be able to inhibit SD sequence evolution due to mRNA structural constraints (Bentele *et al.*, 2013; Goodman *et al.*, 2013; Kosuri *et al.*, 2013; Mutalik *et al.*, 2013; Espah Borujeni *et al.*, 2014).

The efficiency of the three sub processes of translation, namely: initiation, elongation, and termination, greatly depends on efficient initiation, which is generally regarded as the rate-limiting step of the translation process (Liljenstrom *et al.*, 1987; Bulmer, 1991; Xia, 1998; Xia *et al.*, 2007; Prabhakaran *et al.*, 2015). Initiation proceeds at a rate which is dependent on the sequences in the 5′ UTRs of mRNAs (Jacques *et al.*, 1990; Nakagawa *et al.*, 2010).

Translation plays a crucial role in biosynthesis and microbial species typically evolve features to improve translation efficiency. Codon usage in *E. coli*, *Salmonella typhimurium* and *Saccharomyces cerevisiae* greatly is influenced by the availability of their cognate tRNA species (Ikemura, 1981a, 1981b, 1982; Xia, 1998; Prabhakaran *et al.*, 2015), especially in highly expressed genes (Comeron *et al.*, 1998; Duret *et al.*, 1999; Coghlan *et al.*, 2000; Xia, 2007; Prabhakaran *et al.*, 2015). Translation elongation efficiency and evolution of codon usage have been found to be influenced by translation initiation efficiency (Xia *et al.*, 2007; Supek & Smuc, 2010; Tuller *et al.*, 2010; Prabhakaran *et al.*, 2015).

It has been revealed that gene expression can be controlled by three effective determinants equally at different steps of translation process: (1) properties local to the 5' end of the mRNA transcript such as the SD sequence (Hui & de Boer, 1987; de Smit & van Duin, 1994; Olsthoorn *et al.*, 1995), and start codon (Hartz *et al.*, 1991; Ringquist *et al.*, 1992; Ma *et al.*, 2002) influence translation efficiency at the level of initiation; (2) when translation initiation is optimal, codon-anticodon adaptation (Ikemura, 1981b; Xia, 1998) at the translation elongation step becomes rate limiting; (3) if the previous two steps are controlled for, the termination step will limit the efficiency of translation via the effectiveness of the stop codon (Wei, Wang, Xia 2016) and its flanking signals (M. Li *et al.*, 2003).

Interestingly, McLaughlin and colleagues postulated that *Bacillus subtilis* requires more stringent base-pairing of the SD sequence to the 3' end of 16S rRNA, relative to *E. coli*, in order to form an efficient translation initiation complex (McLaughlin *et al.*, 1981). There is only a minor difference between the sequences of the 3' ends of *B. subtilis* and *E. coli* 16S rRNA (Gold *et al.*, 1981; Murray *et al.*, 1982), which suggests that the SD sequence of most *E. coli* mRNAs should base-pair fairly easily with *B. subtilis* 16S rRNA; however, *B. subtilis* requires more accurate base-pairing between the SD sequences and 16S rRNA to initiate translation than does *E. coli*, so this may not always hold true. This could explain the translational discrimination seen against most *E. coli* genes (Band & Henner, 1984).

Our study stems from the findings on these more stringent SD requirements for mRNA processing in gram-positive *B. subtilis* (Band & Henner, 1984). Thus, SD sequences which are appropriate for gene expression in *E. coli* might not be efficient in *B. subtilis* (Band & Henner, 1984) whereas gram-negative *E. coli* are expected to be more flexible in their SD requirements for mRNA processing.

We hypothesize that differences in the prevalence of SD motifs between *B. subtilis* and *E. coli* arise as a result of changes in the free 3' end of 16S rRNA which may have led *B. subtilis* and *E. coli* to evolve differently. Furthermore, these differences may be exacerbated by the absence of RPS1 in *B. subtilis*. We predict that, based on changes in free 3' end of 16S RNA, we will observe some SD sequence motifs that are more prevalent in terms of their proportion and will work better in either *E. coli* or *B. subtilis* than in the other. Furthermore, *E. coli* is expected to be more amenable to the acquisition of SD motifs that do not perfectly correspond with its free 3' 16S rRNA end than *B. subtilis*.

2.3 Material and Methods

2.3.1 Retrieval of genome sequence data

The annotated whole genome sequences for *Escherichia coli K12* (Accession # NC_000913.3) and *Bacillus subtilis 168* (Accession # NC_000964.3) in GenBank format were downloaded from the National Center for Biotechnology Information (NCBI) database (http://www .ncbi.nlm.nih.gov) and were used in subsequent analyses.

2.3.2 Designation of highly and lowly expressed genes

Genes were delimited as HEGs or LEGs on the basis of two metrics: steady state protein abundance levels taken from PaxDB, and I_{TE} scores. I_{TE} scores were computed with DAMBE using the default .I_{TE} files for *E. coli* and *B. subtilis* that were included in the distribution. DAMBE's I_{TE} function has four settings that differ in their treatment of synonymous codon families and we selected the option breaking six-fold degenerate codon families into four and two-fold families. For *E. coli* and *B. subtilis*, the top and bottom 10% of genes for both of these metrics were designated as HEGs and LEGs, respectively.

2.3.3 Identification of Shine-Dalgarno Sequences

In order to identify the free 3′ ends of the 16S rRNAs for *B. subtilis* and *E. coli*, we examined their corresponding 16S rRNA secondary structure schematics from the Comparative RNA Web Site & Project (http://www.rna.icmb.utexas.edu) which is curated by the Gutell Lab at the University of Texas at Austin. The schematics include base pairing interactions that are predicted based on the minimum free energy (MFE) state of the structure as well as those that have been determined empirically. For both species, we defined the end containing the aSD used in subsequent analyses as the stretch of single-stranded terminal 3′ nucleotides which were experimentally confirmed previously (Shine & Dalgarno, 1974; Brosius *et al.*, 1978; Gold *et al.*, 1981; Luhrmann *et al.*, 1981; Murray & Rabinowitz, 1982; Band & Henner, 1984; Tu *et al.*, 2009). The structure of the free 3’ ends is included in Figure 2.2. To verify consistency, we also compared these sequences with all annotated copies of the 16S rRNA in the genome (Starmer *et al.*, 2006).

In this manner, the sequence of the 3' 16S rRNA end used in our analysis for *E. coli* is 3'-AUUCCUCCACUAG-5' (Shine & Dalgarno, 1974; Brosius *et al.*, 1978; Gold *et al.*, 1981; Luhrmann *et al.*, 1981; Band & Henner, 1984; Tu *et al.*, 2009) because, based on the E. coli SSU rRNA secondary structure (Woese *et al.*, 1980; Noah *et al.*, 2000; Yassin *et al.*, 2005; Kitahara *et al.*, 2012; Prabhakaran *et al.*, 2015), these are the 13 nt at the 3’ end of the 16S rRNA that are free to base-pair with the SD sequence (hereafter referred to as the 16S tail). Likewise, the sequence of the 16S tail used in our analysis for *B. subtilis* is 3'-UCUUUCCUCCACUAG-5' (Murray & Rabinowitz, 1982; Band & Henner, 1984). This is similarly based on the *B. subtilis* SSU rRNA secondary structure (Noah *et al.*, 2000) in which there are 15 nt (an additional 2 free nt relative to *E. coli*) in the 16S tail.

Since the SD sequence is characterized by its ability to interact with the free 3' end of the 16S rRNA, the single-stranded segments determined above were used to identify putative SD sequences using the Data Analysis in Molecular Evolution and Biology (DAMBE) integrated software package (Xia, 2013). This was accomplished by extracting the 30 nt preceding the initiation codon of each annotated gene for both genomes and then looking for hits ranging from 4-12 consecutive nucleotides long between these upstream sequences and their respective 3' 16S rRNA free ends (i.e. we searched for Watson-Crick base-pairings between the *E. coli* upstream sequences and *E. coli* 16S tail, and did the same for *B. subtilis*). In addition to this search, we also performed a reciprocal search in which Watson-Crick base-pairings were detected between the opposite 16S tail sequences and upstream sequences (e.g. base-pairings between the *E. coli* 16S tail and the *B. subtilis* upstream sequences). This step was essential in determining the presence and frequency of SD motifs that are expected to be detected in only one of the two species due to differences in their 16S tails. DAMBE outputs several metrics for each hit which are useful in subsequent analysis.

2.3.4 Analysis of Putative Shine-Dalgarno Sequences

Based on the DAMBE output generated for each hit in the search described above, we used a method introduced by Prabhakaran, Chithambaram, and Xia (2015) in order to identify SD sequences from the potential hits. The distributions shown in Figure 2.1 were used to represent the constraint imposed by the SD-aSD interaction. Thus, all hits that did not optimally position the ribosome near the start codon were excluded from analysis. These optimal ranges included 10-22 nt for *E. coli* and 11-22 nt for *B. subtilis*. From the remaining pool of hits, those matching the motifs that we expected to be highly underrepresented in either species were totaled and divided by the sum of all detected motifs within the specified ranges in order to determine the proportion

of the population they represent. The proportion of SD motifs for each species was calculated by the total number of SD counts within the optimal range divided by the total number of genes for a given species.

2.4 Results and discussion

In order to assess SD motifs that have arisen as a result of the differential evolution between the free ends of the 16S rRNAs in these species, we compared the 30 nt preceding the start codon for all genes in each species with both 3' TAILs to identify the proportions of genes containing hits for SD_{Ec} and SD_{Bs}. We defined sets of HEGs and LEGs for each species on the basis of Index of Translation Elongation (I_{TE}) scores (Xia, 2015) computed using (Xia, 2013) and steady state protein abundance values obtained from PaxDB (Wang *et al.*, 2012).

2.4.1 *B. subtilis* genes exhibit a stronger preference for SDBs than *E. coli* genes for SDEc

When the proportions of all SD_{Bs} motifs examined (AAAG, AGAA, and GAAA beginning motifs), about 30% of the genes in species are accounted for (Table 2.2). Strikingly, the reliance of *B. subtilis* on SD_{Bs} motifs is about three-fold higher than what is observed for the usage of SD_{Ec} motifs in *E. coli* (Table 3). In keeping with this observation, *B. subtilis* genes exhibit a stronger avoidance of SD_{Ec} (Table 3) motifs than does *E. coli* for SD_{Bs} (Table 2.2) as the proportion of genes which contain SD_{Ec} in *B. subtilis* is essentially half that of *E. coli* genes containing SD_{Bs}. In fact, *E. coli* appears to contain a slightly higher proportion of genes with SD_{Bs} than those with SD_{Ec}.

These findings corroborate earlier experimental evidence (McLaughlin *et al.*, 1981; Band & Henner, 1984) which supports the notion that *B. subtilis* requires a more stringent Shine-

Dalgarno region for gene expression than does *E. coli*. Band and Henner used a series of plasmids which differed in sequence at the SD region preceding the leukocyte interferon-A gene in order to investigate the efficiency of translation initiation in both *E. coli* and *B. subtilis*. They hypothesized that a SD sequence which is sufficient for gene expression in *E. coli* might not be functional in *B. subtilis*. One of their key observations was that changes in the SD sequences of *B. subtilis* must lead to more stringent base pairing with the 3' TAIL, relative to the changes in *E. coli*, in order to be effective in translation initiation. This idea was also supported by the work of McLaughlin and colleagues (1981). Furthermore, the authors note it is not merely that SD sequences in *B. subtilis* require more stringent complementarity with the 3' TAIL than is observed for *E. coli* that is of importance, but also that the nature of the sequences flanking the translation initiation region (TIR) influence the efficiency of translation initiation.

It is important to consider the nature of TIR flanking sequences because they can potentially generate secondary structure which may interfere with ribosomal binding and inhibit start codon accessibility (Ringquist *et al.*, 1992). This is particularly relevant in *B. subtilis*, which lacks a fully functional RPS1 and may be unable to effectively deal with strong secondary structure in the TIR (Vellanoweth *et al.*, 1992). Since *E. coli* does have a functional S1 protein, this may allow it to compensate for less stringency in the SD-aSD interaction with the ability to effectively reduce secondary structure that would otherwise embed the TIR (Roberts *et al.*, 1989; Farwell *et al.*, 1992; Tzareva *et al.*, 1994). Although the stringency of the SD-aSD interaction is important to consider, the strength of such an interaction is also an important factor as an SD that does not bind to mRNA strongly enough will not confer any specificity with respect to localizing the ribosomal P site at the start codon (Farwell *et al.*, 1992; Komarova *et al.*, 2002; Duval *et al.*, 2013). Conversely, if the SD-aSD interaction is too strong, this may lead to excessive translational

pausing at SD-like sequences which will hinder that rate of protein production (G. W. Li *et al.*, 2012).

2.4.2 SD motifs are weaker in *E. coli* than in *B. subtilis*

Across all motifs analyzed in both species, those observed in *E. coli* tend to be shorter than those in *B. subtilis*, especially in cases where base-pairing in the SD-aSD interaction is semi-specific implying less stability. Generally, the proportions of all SD motifs observed in *B. subtilis* are relatively similar over the range of 5-8 nucleotides in length (Table 2.2, 2.3), suggesting that SD strength may be more fluid between the two species than we initially predicted. In all cases for *B. subtilis*, a sharp decline in motif frequency was observed beyond a length of 8 nt, regardless of base pairing specificity. This generally observed higher tolerance for longer motifs in *B. subtilis* may suggest that SD strength, which is proportional to the length of the motif, plays a larger role in gene regulation at the level of translation in *B. subtilis* than in *E. coli* when imperfect base pairing in the SD-aSD interaction is considered.

The SD_{Ec} motifs in *E. coli* follow a similar trend to those of *B. subtilis*, except that their proportions are similar over the range of 4-7 nt in length and the abrupt decline in frequency occurs after this point. Notably, this length range is consistent with that which has been traditionally defined in *E. coli* (Shine & Dalgarno, 1974; Ringquist *et al.*, 1992; Chen *et al.*, 1994; Zheng *et al.*, 2011). Conversely, the proportions for many of the SD_{Bs} sequences considered in *E. coli* do not exceed 6 nt in length. In fact, more than half of the genes considered in this context have SD sequences that are only 4-5 nt long. Despite the fact that many observed SD sequences in *E. coli* are short, this is not necessarily an indication of effectiveness. Ringquist and others noted in their 1992 study that having UAAGGAGG as an SD sequence enables four-fold higher protein production than does AAGGA. In this context, it makes sense that many of the genes observed do

not use such a potent SD sequence because, in practice, there are relatively few genes that require very high expression to ensure cellular proliferation. Resultantly, many genes are inducible rather than constitutive and are only active during niche scenarios in order to prevent the cell from wasting resources that could better be directed towards more fundamental processes. This notion is concordant with our observations for SD_{Ec} motifs in *E. coli* HEGs wherein more than half of the observed SDs are greater than 6 nt in length (Table 4).

For the vast majority of *E. coli* and *B. subtilis* mRNAs, the recognition of the translation initiation codon is mainly through the match between the SD sequences and the aSD sequences (Shine & Dalgarno, 1974; Steitz & Jakes, 1975; Hui & de Boer, 1987; Jacob *et al.*, 1987; Komarova *et al.*, 2002; Vimberg *et al.*, 2007). It has been reported that modifying the SD or aSD to disrupt base pairing will reduce protein production (Hui & de Boer, 1987). Likewise, mutating aSD to restore the pairing restores the protein production. The canonical aSD sequence on the 16S rRNA 3' tail is often cited as 5'-ACCUCCU-3' in *E. coli* (Shine & Dalgarno, 1974), but its main core has been recognized as CCUCC, within the overwhelming majority of surveyed prokaryotes (Ma *et al.*, 2002; Nakagawa *et al.*, 2010; Lim *et al.*, 2012), although there are a few exceptions. Nevertheless, diversity of SD usage has been observed between different organisms, often based on variation in sequences outside of the core CCUCC motif. Moreover, not all the genes within a given species necessarily initiate via SD-aSD mechanisms; thus, such a pairing is not always crucial in determining the correct initiation codon (Calogero *et al.*, 1988; Melancon *et al.*, 1990; Fargo *et al.*, 1998). For example, leaderless messengers (Shean *et al.*, 1992; C. J. Wu & Janssen, 1997; Van Etten *et al.*, 1998), mRNAs bearing plant viral leader sequences (Wilson, 1986; Gallie *et al.*, 1989; Tzareva *et al.*, 1994), *Chlamydomonas reinhardtii* chloroplast mRNAs (Fargo *et al.*, 1998), and the *tuf* mRNA of *Mycoplasma genitalium* (Loechel *et al.*, 1991), are all recognized and

translated by *E. coli* ribosomes although they lack any SD-like sequence upstream of the start codon. Such leaderless genes often have an AUG start codon in *E. coli* and are known to proceed efficiently through translation despite lacking an SD motif (O'Donnell *et al.*, 2002; Krishnan *et al.*, 2010; Vesper *et al.*, 2011; Giliberti *et al.*, 2012; Prabhakaran *et al.*, 2015) or in the halophilic archaeon *Halobacterium salinarum* (Sartorius-Neef *et al.*, 2004).

2.4.3 *E. coli* HEGs exhibit stronger selection for SD_{Ec} than LEGs

Based on the proportions of SD_{Ec} and SD_{Bs} observed in the HEGs and LEGs of *E. coli*, there is sufficient evidence to suggest that there is selection mediated by its 3' TAIL. There is a three-fold higher usage of SD_{Ec} in HEGs relative to LEGs for *E. coli* with an SD length ranging from 4-7 nt (Table 4) which is consistent with values reported in the literature(Shine & Dalgarno, 1974; Ringquist *et al.*, 1992; Chen *et al.*, 1994; Zheng *et al.*, 2011). There is also no indication of selection for the use of SD_{Bs} in *E. coli* as the proportions of genes containing them are consistently low regardless of expression level (Table 5). These trends are consistent with our predictions and indicate that 3' TAIL-mediated selection in *E. coli* has a direct impact on the efficiency of protein biosynthesis. It is notable that the weak preference for SD_{Ec} observed overall for *E. coli* genes emphasizes their importance in effective protein production given that selection for SD_{Ec} in *E. coli* HEGs is high. This is because HEGs encompass only a small subset of genes in a given organism, therefore the weak trend for SD_{Ec} overall in *E. coli* is predominantly due to the strong selection for SD_{Ec} in HEGs. Interestingly, this pattern is inverse in *B. subtilis*.

2.4.4 LEGs in *B. subtilis* more strongly prefer SD_{Bs} than HEGs

Although *B. subtilis* has almost twice the proportion of SD_{Ec} present in its HEGs than in its LEGs, this value is still lower than any of its proportions for SD_{Bs} motifs. Furthermore, this value

is also lower than the proportion of SD_{Ec} found in the LEGs of *E. coli*, thus it is not suggestive of selection for SD_{Ec} in *B. subtilis* which is consistent with our expectations. Conversely, the proportion of SD_{Bs} in *B. subtilis* LEGs is almost twice that which is seen in the HEGs. One limitation in our analysis that may account, at least in part, for this observation is our assumption that every gene is preceeded by at least 30 nt at the RNA level. In practice, there are a subset of genes that have less than 30 nt preceeding the start codon, and some genes have no 5' untranslated region (UTR) at all. Despite this, the limitation is unlikely to skew the data in any particular direction since there is no selection acting to preserve a functional SD motif at the DNA level.

It is plausible that the 3' end maturation process which generates the functional 16S rRNA may introduce more heterogeneity in *B. subtilis* than in *E. coli*. Similarly to 5' end maturation events, 3' processing first occurs through an endonucleolytic cleavage liberating the 16S rRNA end (Britton *et al.*, 2007; Kurata *et al.*, 2015), this is followed by exonucleases nibbling away excess nucleotides in a 3' to 5' fashion to generate the mature end (Yao *et al.*, 2007). It is common for this 3'-5' exonuclease activity to introduce some degree of heterogeneity in the mature products and it is concievable that degree of this heterogeneity may differ between the distantly related *E. coli* and *B. subtilis*, causing the end of *E. coli* to more consistently reflect the end used in our analysis whereas *B. subtilis* may occasionally contain fewer nucleotides in its 3' TAIL. This would account for the absence of strong selection for SD_{Bs} in HEGs mediated by the unique nucleotides in the 3' TAIL of *B. subtilis* (Table 5) since such variability in the 3' TAIL could potentially impact effective protein synthesis in these genes. Furthermore, this notion is compatible with our observations given that selection for SD_{Bs} in *B. subtilis* LEGs is strongest in motifs beginning with AAAG (Table 5) which requires only one unique nucleotide relative to the *E. coli* 3' TAIL.

The problem of recognizing the translation initiation codon has been solved differently within prokaryotes (Komarova *et al.*, 2002). In prokaryotes, ribosomes are responsible for the differentiation of the initiation codon (initiator AUG) from synonymous AUG triplets or non-AUG codons throughout mRNA via specific signals in the vicinity of translation start site (Schneider *et al.*, 1986; Dreyfus, 1988; Gold, 1988; Gualerzi *et al.*, 1990; Ringquist *et al.*, 1992; Komarova *et al.*, 2002).

For the vast majority of *E.coli* and bacteriophage mRNAs, the recognition of the translation initiation codon is mainly through the match between the SD sequences about 10 nt upstream of the translation initiation codon and the aSD sequences at the 3' end of the small ribosomal rRNA (Shine & Dalgarno, 1974; Steitz & Jakes, 1975; Hui & de Boer, 1987; Jacob *et al.*, 1987; Komarova *et al.*, 2002; Vimberg *et al.*, 2007). aSD sequence on their 3' tails is highly conserved, but unique across prokaryotes. Since the discovery of aSD in *Escherichia coli* (Shine & Dalgarno, 1974), as 5'-ACCUCCU-3', its main core has been recognized as 5'-CCUCC-3', within the overwhelming majority of surveyed prokaryotes (Ma *et al.*, 2002; Nakagawa *et al.*, 2010; Lim *et al.*, 2012), although there are a few exceptions. Nevertheless, a huge heterogeneity of SD usage has been seen between different organisms and not all the genes within the whole genome of a given species initiate via SD-aSD mechanisms. Interestingly, such pairing is not always crucial in determining the correct initiation codon (Calogero *et al.*, 1988; Melancon *et al.*, 1990; Fargo *et al.*, 1998), and presence of functional mRNAs entirely lacking SD signposts that prokaryotic ribosomes have other mechanisms for start site selection. For example, leaderless messengers (Shean & Gottesman, 1992; C. J. Wu & Janssen, 1997; Van Etten & Janssen, 1998), mRNAs bearing plant viral leader sequences (Wilson, 1986; Gallie & Kado, 1989; Tzareva *et al.*, 1994), *Chlamydomonas reinhardtii* chloroplast mRNAs (Fargo *et al.*, 1998), and the tuf mRNA of

Mycoplasma genitalium (Loechel *et al.*, 1991), are all recognized and translated by *E.coli* ribosomes although they lack any SD-like sequence upstream of the start codon. For instance, there is a subset of leaderless genes with an AUG start codon in *E.coli* that are known to proceed efficiently through translation despite lacking an SD motif (O'Donnell & Janssen, 2002; Krishnan *et al.*, 2010; Vesper *et al.*, 2011; Giliberti *et al.*, 2012; Prabhakaran *et al.*, 2015) or in the halophilic archaeon *Halobacterium salinarum* (Sartorius-Neef & Pfeifer, 2004).

The most energetically stable SD-aSD interaction indices were identified by significant studies that determined the base pairing potential of SD-aSD duplex using a free energy approach (Schurr *et al.*, 1993; Osada *et al.*, 1999; Starmer *et al.*, 2006). SD motif length and content are the two degrees of freedom that determine the base pairing potential between SD and aSD.SD motif can base pair either partially or completely with an aSD sequence; accordingly, different variations of SD sequence exists. Ringquist and his colleagues examined gene expression in *E. coli* through varying the length of the SD sequence (1992). They noticed that SD motif UAAGGAGG with eight bases roughly enables four fold higher gene expression than the SD motif AAGGA with five bases (Ringquist *et al.*, 1992). A long SD-aSD duplex has been verified to inhibit translation because of a strong interaction between 30S subunit and RBS. Therefore, overly strong binding with SD can prevent the ribosome from effectively proceeding to elongation (Komarova *et al.*, 2002).

The protein abundance among different species within a single cell can differ by several orders of magnitude and several points of control are critical for regulation of the expression of individual proteins (Dekel & Alon, 2005; Kudla *et al.*, 2009; Salis *et al.*, 2009; Y. Taniguchi *et al.*, 2010). The first stage in the pathway of gene expression is the transcription of the gene of interest; however, transcription is insufficient to ensure protein expression. Research that was done in

different organisms has demonstrated that mRNA abundances only modestly predict protein abundances (Lu *et al.*, 2007; Y. Taniguchi *et al.*, 2010; Vogel *et al.*, 2010; Schwanhausser *et al.*, 2011; Vogel *et al.*, 2012).

It has been reported by (Hui & de Boer, 1987) that modifying the SD or aSD to disrupt base pairing will reduce protein production. Likewise, mutating aSD to restore the pairing restores the protein production. However, numerous studies have provided evidences suggesting that the SD is not needed for translation initiation such as 1) the classic Nirenberg and Matthaei experiment with poly-U, 2) the work that was done in 1990 by P. Melancon and his colleagues. He attested that the removal of the last ~30 nucleotides in 16 rRNA led to reduce protein production, but translation initiation was still at the same initiation codon. The main concern was that the removal of the 30 nucleotides was creating another unpaired 3' tail that can form base pairs with the 5' UTR or mRNA; 3) another work which was done by Fargo and his colleagues in 1998 convey the analysis of chloroplast transformants of C. reinhardtii and transformants of E. coli including the wild-type and mutant reporter constructs which revealed that mutagenic replacement of the putative SD sequences had no influence on the expression of either the reporter genes used in experiment. It was suggested that Chloroplast transformants with the canonical SD sequence also indicated no differences in reporter gene expression, while expression of the reporter genes was increased by 10 to 30% in the *E. coli* transformants. The final results suggest that even though SD-dependent initiation predominates in *E. coli*, this bacterium also has the capability to initiate translation by an SD-independent mechanism. On the other hand, plant chloroplasts seem to have adopted the SD-independent mechanism for translational initiation of most mRNAs.

2.5 Conclusion

In summation, our findings suggest that *B. subtilis* is more reliant on stringent base pairing in the SD-aSD interaction and tends to prefer longer SD motifs than *E. coli* regardless of gene expression level. Despite *E. coli* exhibiting a weaker overall preference for SD_{Ec} than *B. subtilis* for SD_{Bs} in genes, the trend is largely associated with the subset of HEGs in *E. coli* which, combined *E. coli*'s low usage of SD_{Bs} motifs, implies that 3' TAIL-mediated selection is a driving force in protein production for this species. Furthermore, the usage of SD_{Bs} in *B. subtilis* is better accounted for by LEGs than HEGs which may imply a higher degree of heterogeneity in the *B. subtilis* 3' TAIL. Moving forward, it would be interesting to investigate the effects of 3' TAIL-mediated selection in other species to determine whether or not more of them rely on sequence differences, as opposed to conserved regions, between other organisms to achieve varying levels of expression.

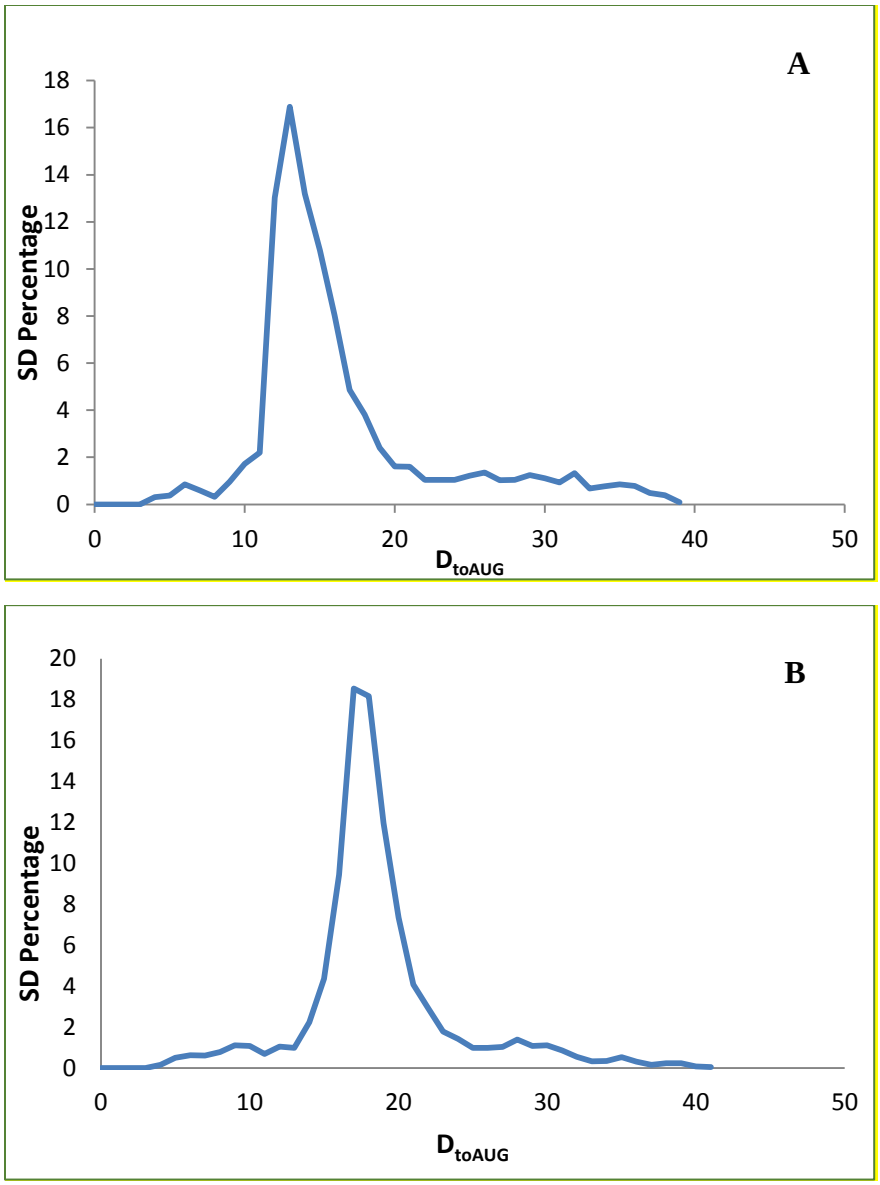

Figure 2.1. The distribution of matching hits between the putative SD and the 16S rRNA characterized by the relative position of the ribosome to the start codon for (A) *E. coli* and (B) *B. subtilis*. The y-axis represents the percentage of SD motifs and x-axis represents the distance to AUG codons (D_{toAUG}) from the 3'end of 16S rRNA to start codon and the upstream 30 nucleotides of CDSs.

Table 2.1. The 3’ end of 16S rRNA of *E.coli* and *B.subtilis* which are free to base pair with Shine dalgarno sequence.

Species	16S rRNA sequence*	SD motifs		
E.coli	3'-**A**U**U**CCUCCACUAG-5'	**UAAG**,UAAGG,UAAGGA,UAAGGAG,UAAGGAGG, UAAGGAGGUG		
B.subtilis	3'-**UCU**U**U**CCUCCACUAG-5'	**AGAA**	**GAAA**	**AAAG**
		AGAAA	**GAAAG**	AAAGG
		AGAAAG	GAAAGG	AAAGGA
		AGAAAGG	GAAAGGA	AAAGGAG
		AGAAAGGA	GAAAGGAG	AAAGGAGG
		AGAAAGGAG	GAAAGGAGG	AAAGGAGGU
		AGAAAGGAGG	GAAAGGAGGU	AAAGGAGGUG
		AGAAAGGAGGU	GAAAGGAGGUG	AAAGGAGGUGA
			GAAAGGAGGUGA	

* - 1 - Highlighted area shows the differences in the base composition between two species. 2 - The SD motifs shown are derived from differences in free 3’ tail of 16S rRNA in both species. Only motifs observed in our analysis are shown.

Table 2.2 SD_{Bs} hits in all *Bacillus subtilis* and *Escherichia coli* genes.

SD_{Bs} motifs	Occurrence in *B. subtilis*		Occurrence in *E.coli*	
	Count	Proportion	Count	Proportion
AGAA	12	0.0029	51	0.0123
AGAAA	66	0.0158	60	0.0145
AGAAAG	60	0.0144	14	0.0034
AGAAAGG	54	0.0129	7	0.0017
AGAAAGGA	60	0.0144	6	0.0014
AGAAAGGAG	28	0.0067	4	0.0010
AGAAAGGAGG	11	0.0026	1	0.0002
AGAAAGGAGGU	1	0.0002	0	0
Subtotal	292	0.0699	143	0.0345
GAAA	16	0.0038	65	0.0157
GAAAG	41	0.0098	28	0.0068
GAAAGG	68	0.0163	18	0.0043
GAAAGGA	51	0.0122	15	0.0036
GAAAGGAG	57	0.0137	10	0.0024
GAAAGGAGG	18	0.0043	1	0.0002
GAAAGGAGGU	3	0.0007	0	0
GAAAGGAGGUG	1	0.0002	0	0
GAAAGGAGGUGA	1	0.0002	0	0
Subtotal	240	0.0575	137	0.0331
AAAG	19	0.0046	38	0.0092
AAAGG	171	0.0410	83	0.0200
AAAGGA	76	0.0182	101	0.0244
AAAGGAG	222	0.0532	64	0.0155
AAAGGAGG	143	0.0343	6	0.0014
AAAGGAGGU	31	0.0074	3	0.0007

AAAGGAGGUG	6	0.0014	0	0
AAAGGAGGUGA	3	0.0007	1	0.0002
Subtotal	671	0.1607	296	0.0715
Total	1203	0.2881	576	0.1391

Table 2.3. SD_{Ec} hits in all *Bacillus subtilis* and *Escherichia coli* genes.

SD_{Ec} motifs	Occurrence in *E.coli*		Occurrence in *B. subtilis*	
	Count	Proportion	Count	Proportion
UAAG	85	0.0205	15	0.0036
UAAGG	91	0.0220	54	0.0129
UAAGGA	151	0.0365	30	0.0072
UAAGGAG	117	0.0283	74	0.0177
UAAGGAGG	10	0.0024	74	0.0177
UAAGGAGGU	0	0	14	0.0033
UAAGGAGGUG	1	0.0002	6	0.0014
Total	455	0.1099	267	0.0640

Table2.4. SD_{Ec} hits in all highly and lowly expressed genes.

SD_{Ec} motifs	**Occurrence in *E. coli***				**Occurrence in *B. subtilis***			
	HEGs		LEGs		HEGs		LEGs	
	Count	Proportion	Count	Proportion	Count	Proportion	Count	Proportion
UAAG	22	0.0053	7	0.0017	1	0.0002	3	0.0007
UAAGG	32	0.0077	6	0.0014	4	0.0010	3	0.0007
UAAGGA	36	0.0087	20	0.0048	3	0.0007	0	0
UAAGGAG	40	0.0097	12	0.0029	9	0.0022	10	0.0024
UAAGGAGG	2	0.0005	1	0.0002	14	0.0034	2	0.0005
UAAGGAGGU	0	0	0	0	0	0	1	0.0002
UAAGGAGGUG	0	0	0	0	4	0.0010	0	0
Total	132	0.0319	46	0.0111	35	0.0084	19	0.0046

Table 2.5. SD_{Bs} hits in all highly and lowly expressed genes.

SD_{Bs} motifs	Occurrence in *B. subtilis*				Occurrence in *E. coli*			
	HEGs		LEGs		HEGs		LEGs	
	Count	Proportion	Count	Proportion	Count	Proportion	Count	Proportion
AGAA	0	0	2	0.0005	3	0.0007	3	0.0007
AGAAA	2	0.0005	8	0.0019	7	0.0017	9	0.0022
AGAAAG	6	0.0014	4	0.0010	1	0.0002	1	0.0002
AGAAAGG	3	0.0007	6	0.0014	1	0.0002	0	0
AGAAAGGA	4	0.0010	2	0.0005	2	0.0005	0	0
AGAAAGGAG	2	0.0005	3	0.0007	1	0.0002	0	0
AGAAAGGAGG	1	0.0002	2	0.0005	0	0	0	0
AGAAAGGAGGU	0	0	0	0	0	0	0	0
Subtotal	18	0.0043	27	0.0065	15	0.0036	13	0.0031
GAAA	0	0	2	0.0005	5	0.0012	10	0.0024
GAAAG	2	0.0005	7	0.0017	3	0.0007	1	0.0002
GAAAGG	3	0.0007	11	0.0026	0	0	0	0
GAAAGGA	4	0.0010	5	0.0012	5	0.0012	0	0
GAAAGGAG	2	0.0005	6	0.0014	1	0.0002	1	0.0002
GAAAGGAGG	2	0.0005	2	0.0005	0	0	0	0
GAAAGGAGGU	0	0	0	0	0	0	0	0
GAAAGGAGGUG	0	0	0	0	0	0	0	0
GAAAGGAGGUGA	0	0	0	0	0	0	0	0
Subtotal	13	0.0031	33	0.0074	14	0.0034	12	0.0029
AAAG	1	0.0002	4	0.0010	2	0.0005	2	0.0005
AAAGG	8	0.0019	20	0.0048	7	0.0017	12	0.0029
AAAGGA	5	0.0012	10	0.0024	10	0.0024	9	0.0022
AAAGGAG	17	0.0041	26	0.0062	7	0.0017	7	0.0017
AAAGGAGG	14	0.0033	21	0.0050	1	0.0002	0	0
AAAGGAGGU	2	0.0005	1	0.0002	1	0.0002	0	0
AAAGGAGGUG	1	0.0002	0	0	0	0	0	0
AAAGGAGGUGA	0	0	0	0	0	0	1	0.0002
Subtotal	48	0.0115	82	0.0196	28	0.0068	31	0.0075
Total	79	0.0189	142	0.0335	57	0.0138	56	0.0135

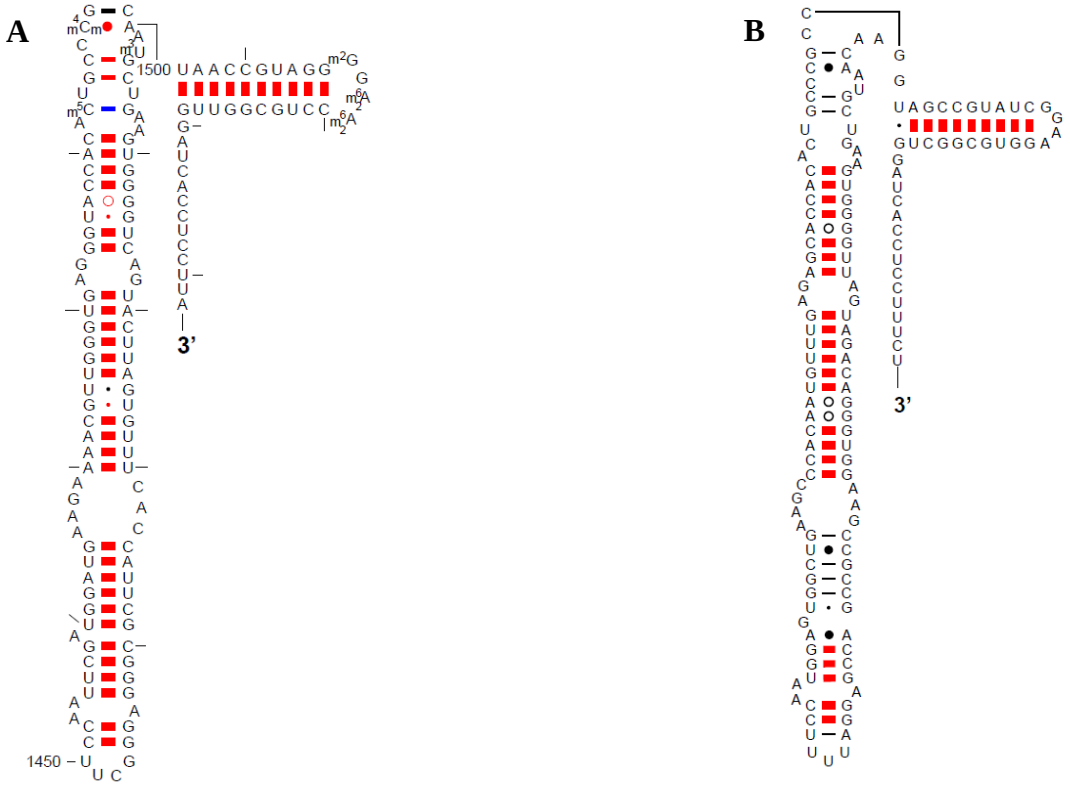

Figure 2.2. The free 3' end of SSU rRNA are shown in both species. (A) Hypothetical secondary structure of *E. coli* 16S rRNA. (B) Hypothetical secondary structure of *B. subtilis* 16S RNA. Citation and related information available at http://www.rna.icmb.utexas.edu

Chapter 3

Apparent relationship of SD incidence and differential codon adaptation in bacteria and archaea

3.1. Abstract

In prokaryotes, translation initiation generally requires SD binding and weak secondary structure around initiation codon. Recognition of initiation codon proceeds through four categories: (i) presence of SD sequence alone, (ii) presence of SD sequence aided by S1, (iii) absence of SD sequences but aided by S1 as seen in leaderless open reading frames (ORFs) in bacteria, and (iv) absence of SD and S1 protein as seen in most archaeal genes. Although translation initiation efficiency depends on codon usage differences across prokaryotic genomes, among genes from a single prokaryotic species, it differs based on SD base-pairing strength and location. From these observations, we test the hypothesis that translation initiation efficiency varies across the genes within the genomes of forty two prokaryotic. We measured translation initiation by: i) the strength and position of SD sequence and ii) the stability of the secondary structure flanking the start codon, which affect accessibility of the start codon.

3.2. Introduction

The key to a cell's survival is protein synthesis, a process which consumes ~75% of available cellular energy (Harold, 1986). Translation involves the decoding of genetic information in the form of messenger RNA (mRNA) into amino acids. The mRNA template is made up of different combinations and amounts of four nucleotides: guanine, adenine, uracil, and cytosine. However, only a combination of three nucleotides codes for an amino acid. Therefore, there are 4^3= 64 codons of which 61 code for 20 amino acids. The codons are synonymous in that more than one

codon can code for the same amino acid but are also specific in that none of them specifies another amino acid (Gouy & Gautier, 1982). Across all species, there exist codon usage biases (CUB)(Codon usage bias refer to differences in the frequency of occurrence of synonymous codons in coding DNA) whereby synonymous codon (Synonymous codons refer to that more than one codon can code for the same amino acid) differ in their frequency of use among different genes (Grantham *et al.*, 1980). Genes, especially highly expressed genes, select codons having abundant transfer RNAs (Ikemura, 1981b; Bennetzen & Hall, 1982; Ikemura, 1982; Xia, 1998; Kanaya *et al.*, 1999; Emery *et al.*, 2011). Selection affects highly expressed genes (HEGs) the most because the translation efficiency of these has the greatest effect on growth (Ehrenberg *et al.*, 1984).

Different mutations and selection pressures are known to cause CUB; genome composition (G+C %) and strand asymmetry in nucleotide are considered the primary factors among mutation pressures (Muto & Osawa, 1987; McInerney, 1998; Xia, 2012). In 1981, Ikemura observed in *E. coli* a close relationship between codon frequency and tRNA population in 26 tRNAs such that there was a preferential codon usage based on tRNA availability, especially for abundant protein species.

He proposed that lower codon optimization such as in amino acid synthesizing genes could be attributed to randomized mutation events and the present sequence is one that is in or has reached an equilibrium state between selective pressure and randomized mutation events (Ikemura, 1981b). High gene expression in prokaryotes can be brought about by various mechanisms: weaker secondary structure at the 5' end of mRNA (Kudla *et al.*, 2009; Gu *et al.*, 2010), high transcription levels from strong promoters (Reisbig *et al.*, 2004; Tegel *et al.*, 2011) , high mRNA stability (Hargrove *et al.*, 1989; Liebhaber, 1997; Cheadle *et al.*, 2005) and/or efficient translation initiation by optimal Shine-Dalgarno (SD) sequences.

A study carried out by Osterman *et al*. (2013) in which *E. coli* mRNA features affecting translation initiation and reinitiation were compared showed that efficient translation initiation was: (i) frequently observed or even predominant in genes with AUG and GUG start codons where GUG had marginally higher initiation efficiency than AUG; (ii) strongly inhibited by stable hairpins at any positions between SD beginning and mRNA coding region, or covering start codon, whereas hairpins surrounding the TIR from both 5' and 3' ends stimulated expression; (iii) optimal for either the strongest or moderately strong mRNA when spacer length was considered. Meaning that mRNA with 8 nucleotides and 7 spacer length (8/7) had higher reported expression than 6/7 mRNA, 4/7 and 2/7 mRNA. Similarly, 8/13 mRNA had higher reported expression than 6/13, 4/13 and 2/13 mRNA. Moderately strong mRNA such as 6/10 (6 nucleotides and 10 spacer length) has higher reported expression than 8/10, 4/10, and 2/10 and 6/16 higher reporter expression than 8/16, 4/16, and 2/16 (Osterman *et al.*, 2013); and (iv) partially dependent on SD presence in intercistronic genes and was efficient with stimulation of A/U-rich sequences upstream of SD sequences.

The translation initiation region consists of the initiation codon, the Shine-Dalgarno sequence (absent in some mRNAs), upstream region of SD and downstream region of the initiation codon (Vimberg *et al.*, 2007). The SD sequence is a purine rich region 4-8 nt long, located 5-13 nt (Ringquist *et al.*, 1992; Chen *et al.*, 1994) upstream of the start codon and complementary to the 3'-terminal sequence of the 16S rRNA (Shine & Dalgarno, 1974). SD-aSD interaction is thought to direct the P-site of the 70S ribosome to the initiation codon, initiating translation. The strength of SD-aSD interactions, length of spacer between SD and start codon, and secondary structures flanking the start codon all affect translation initiation. A strong SD-aSD interaction is thought to strengthen the duplex formation between the 16S rRNA and the SD which increases the accuracy

of positioning the start codon at the ribosomal P-site (Kozak, 1999; Osterman *et al.*, 2013). A large 8 nt SD, combined with 7-13 nt produce high protein expression whereas a short 6 nt SD with 7 nt spacer causes inhibition (Osterman *et al.*, 2013). Secondary structures flanking the start codon of polycistronic mRNA can mask the codon, thereby making it inaccessible to the ribosome and inhibiting translation initiation (Kudla *et al.*, 2009; Scharff *et al.*, 2011; Osterman *et al.*, 2013). In addition, effective SD sequences are generally associated with HEGs (Ma et al. 2002).

The TIR also includes A/U-rich regions thought to be sites of ribosomal S1 protein interaction that increase the efficiency of translation initiation (J. R. Zhang *et al.*, 1989; Tzareva *et al.*, 1994; Vimberg *et al.*, 2007). The major difference between the *E. coli* and *B. subtilis* translation machinery is the presence or absence of ribosomal protein S1 respectively (Roberts & Rabinowitz, 1989). S1, encoded by *rpsA*, is an integral part of the translation machinery of all proteobacteria, cyanobacteria, organelles, and many other phylogenetic groups of bacteria, but is non-functional in gram-positive bacteria (Sorokin *et al.*, 1995). The gene does have a homolog in *B. subtilis.* However, the product is truncated and incapable of binding with the ribosome (Sorokin *et al.*, 1995).

In *E. coli*, SD presence is less important for correct initiation than in *B. subtilis*, as many mRNAs such as leaderless mRNAs (Shean & Gottesman, 1992), tuf mRNA from *M. genitalium* (Loechel *et al.*, 1991) and plant viral leaders (Wilson, 1986; Tzareva *et al.*, 1994) lacking SD sequences have been correctly recognized and translated by *E. coli* ribosomes. However, the same is not true for gram-positive *B. subtilis* whose ribosome can only correctly position and translate mRNA with strong SD sequences (SD strength refers to the length of SD sequence and the ability to base pair with aSD sequence) (Sorokin *et al.*, 1995). Recently, Duval and his colleagues in 2013 showed that, in *E. coli*, the first three domains of S1 protein endow the 30S subunit acting as

RNA chaperone activity (RNA chaperone activity open up misfolded RNA structures and do not require ATP) which is essential for binding and unfolding of structured mRNAs, thereby allowing correct positioning of initiation codon for translation. In addition, they also showed that ribosomal protein S1 is not required for all mRNAs; S1 interaction with mRNA having weak SD sequences greatly increases gene expression but it has no effect on expression in mRNA with strong SD sequences.

Archaea, despite a similar lack of nuclei, have a more complex translational apparatus than bacteria (Londei, 2005). Like bacteria, many archaeal mRNAs are polycistronic. However, they display SD sequences only in a minority of known cases. Most archaeal mRNAs either lack SD motifs in the 5'-UTRs or lack 5' UTR entirely (Tolstrup *et al.*, 2000; Slupska *et al.*, 2001; Brenneis *et al.*, 2007). Brenneis and colleagues (2007) characterized the lengths of 5'-UTRs and 3'-UTRs of 40 transcripts from two haloarchaeal species: *Halobacterium salinarium* and *Halofex volcannii* found that less than 10% of all genes in the genome had SD, 5'-UTRs of most leadered transcripts lacked SD, and the majority of transcripts were leaderless. The leaderless translation mechanism is unclear and there are some observations that are subject to debate among researchers. One such observation is mRNAs endowed with 5'-UTRs but lacking SD motifs. SD-led mRNA-ribosome interaction has long been considered essential for prokaryotic translation initiation efficiency. *In vitro* and *in vivo* mutagenesis experiments in Bacteria and Archaea, however, suggest that weakening or disrupting SD motif had adverse effects from little to no protein production (Slupska *et al.*, 2001; Sartorius-Neef & Pfeifer, 2004; Brenneis *et al.*, 2007). Why should SD disruption of mRNA fed to genomes having majority leaderless genes have little or no protein production? Furthermore, while gram-negative bacteria have S1 protein which can promote ribosome binding of leaderless mRNAs within bacterial ORFs, gram-positive (Isono *et al.*, 1976; Sorokin *et al.*,

1995) and archaeal ribosomal S1 proteins do not have the same function as bacterial proteins and the details of their function is not found yet (Lecompte *et al.*, 2002). Therefore, the manner by which the leaderless mRNAs lacking SD motifs are translated is puzzling. A recent study shows that, in *Methanococcus maripaludis*, codon usage is dominated by a large A+T bias thought to be resulting from mutation pressure (Emery & Sharp, 2011). The strength of selected codon usage bias in HEGs of *M. maripaludis* is very close to that of *E. coli* despite a moderately higher growth rate of 2.3h. However, the CUB is restricted to two fold degenerate amino acids (Emery & Sharp, 2011).

Summarizing the key points: i) gram-positive *B. subtilis* have stringent SD requirements for mRNA processing whereas gram-negative *E.coli* are flexible in their SD requirements for mRNA processing, ii) SD sequences are associated more with HEGs than average genes (Ma *et al*. 2002), iii) Archaeal genomes tend to have more leaderless mRNA (Karlin *et al.*, 2005). This leads to the hypothesis that there is differential translation initiation between sixteen archaeal and twenty six bacterial genomes being analyzed and that, within Bacteria and Archaea,

The hypothesis stems from recent findings that have determined that codon adaptation depends on translation initiation efficiency (Supek & Smuc, 2010; Tuller *et al.*, 2010; Prabhakaran *et al.*, 2015; Xia, 2015). When translation initiation is efficient, translation elongation becomes rate limiting and selection pressure increases translation efficiency and drives codon adaptation. However, when translation initiation is inefficient, then selection pressure will not increase translation efficiency because elongation is not rate limiting. Therefore, if translation initiation is more efficient in bacteria than archaea, then a prediction follows that selection for translation elongation will be stronger in bacterial than archaeal genomes.

To test the hypothesis within genes of the bacterial and archaeal genomes, the following predictions are made: i) that gram-positive bacteria have stronger well-positioned SD sequence than gram-negative bacteria, ii) that gram-positive bacteria have weaker secondary structures flanking the start codon, iii) that Euryarchaeota have stronger well-positioned SD-sequence than Crenarchaeota, and iv) that Euryarchaeota have weaker secondary structures flanking the start codon.

3.3. Material and Methods

3.3.1 Genomic data

The GenBank files of 42 prokaryote genomes including: 26 selected from the Protein Abundance Across Organisms database (PaxDB) and the remainder selected for their diversity, were retrieved from the genome database of National Center for Biotechnology Information (NCBI). The selected archaea are phylogenetically divergent and include several representatives of the two kingdoms Euryarcheota and Crenarcheota. The selected bacteria are similarly chosen for their diversity. The 42 genomes include 26 bacteria and 16 archaea species. Selected species name, phylum, accession number and anti-SD are summarized in table 3.1- 3.4.

Pseudogenes and hypothetical protein coding genes were excluded from analysis. Pseudogenes are unable to encode protein and have lost their gene expression, so they are considered to be dysfunctional remnants of genes and most of them contain multiple mutations which cause changes in the reading frame and premature termination (Vanin, 1985).

For tabular and graphical data, the genome names were codified using the first three and the last three letters from the species name, as given in Table 3.9. Gram stain, sporulation nature and habitat were obtained from BACMAP genome atlas (Stothard *et al.*, 2005). The generation

time for each species was obtained from BIONUMBERS database (Milo *et al.*, 2010) which compiles the minimum generation time and optimal temperature for 214 bacterial species. The literature from which the generation time was selected is also provided in the database (Table 3.9).

3.3.2 Identification of Shine Dalgarno sequences

The Shine-Dalgarno sequences were identified by considering the relative position of rRNA 3'-end to the mRNA (Prabhakaran *et al.*, 2015). The distance is from the end of the SSU rRNA to the beginning of the start codon. Thirty nucleotides upstream of the 16S small subunit rRNA were extracted by using the Data Analysis in Molecular Biology and Evolution (DAMBE) software package (Xia, 2013). The extended 16S rRNA were then aligned using DAMBE. The distance from the end of the SSU 16S rRNA to the beginning of the start codon was used to represent distance of Shine Dalgarno sequence to the start codon (Prabhakaran *et al.*, 2015). Using the method created by Prabhakaran, Chithambaram and Xia (2015), we searched upstream 30 nucleotides against the rRNA 3'-end for a match length of at least 4 consecutive bases.

For all the species including *E. coli*, the frequency of SD matches decreased for any distance other than between 10-20nts, similar to the *E. coli* results obtained by Prabhakaran, Chithambaram and Xia (2015). Therefore, we defined SD as a sequence four bases or longer that can pair with rRNA 3'-end to D_{toAUG} within 10-20 nucleotides. For *E.coli*, the distribution of the frequencies of D_{AUG} from 4584 matches peaks at $D_{toAUG} = 13$ and decreases rapidly towards $D_{toAUG} = 10$ and $D_{toAUG} = 22$. An SD such as AGGAG would need six bases between the end of SD and the beginning of AUG to have a $D_{toAUG} = 13$. Similarly, an SD such as AGGAGG would need five bases between the end of SD and the beginning of AUG to have a $D_{toAUG} = 13$ (Prabhakaran *et al.*, 2015).

Although the *E.coli* SSU rRNA secondary structure has 13nt at the 3' end of rRNA (Woese *et al.*, 1980; Yassin *et al.*, 2005), mainly the first six sites are involved in SD-aSD base pairing (Prabhakaran *et al.*, 2015). However, of 756 putative SDs (including 166 GAGGU, 169 AGGU, 154 GUGA, and 267 UGAU), in 30nt upstream, *E.coli* genes have the second A from the 3' end SSU rRNA. This observation is consistent with the observations made by Prabhakaran, Chithambaram and Xia (2015).

After having identified SD sequences for each genomes, we computed three indices for each species: (i) the Index of Translation Elongation (I_{TE}) (Xia 2015), (ii) the proportion of SD-containing genes in highly and lowly expressed genes (P_{SD}-HEG , P_{SD}- LEG), and, (iii) the mean number of consecutively matched sites (M_{SD}). Ma *et al*. (2002) assessed the relationship between SD presence and gene features such as expression level in 30 prokaryote genomes and found a significant positive relationship between SD presence and the predicted gene expression level based on codon usage bias. Additionally, they found that highly expressed genes are more likely than average genes to possess an SD sequence. Furthermore, Xia (2015) developed a new codon adaptation index which takes into account background mutation bias, a feature absent in CAI (by accounting for changes in codons that occur due to the lack of selective pressure such as lowly expressed genes). We therefore used I_{TE} to categorize highly and lowly expressed genes and then determined the proportion of SD containing genes within each category.

3.3.3 Measuring stability of local mRNA secondary structure

Mean folding energy (MFE, kJ/mol) was used as a proxy to measure translation initiation efficiency. The initiation codon accessibility is a key determinant of translation initiation efficiency (Nakamoto, 2006) and the relationship between translation initiation and the secondary structure of sequences flanking the start codon has been researched in *E.coli* (de Smit & van Duin,

1990, 1994; Osterman *et al.*, 2013), *Saccharomyces cerevisiae* (Xia *et al.*, 2011) and other Eukaryotes (Xia *et al.*, 2009). Osterman et al. (2013) found that in *E.coli*, protein production is dramatically reduced when either the SD sequence or start codon is buried by a stable secondary structure. Therefore, we measured the MFE of 40 bases (referred to as MFE_{40nt}) upstream of the start codon where the presence of a hairpin strongly inhibits translation (Osterman *et al.*, 2013). The more negative the MFE value, the greater the stability of the secondary structure. We computed MFE using DAMBE which uses the RNA folding library from Vienna RNA package (Hofacker, 2003). Three criteria: a folding temperature of 37°C, with no lonely pairs and no G/U pairs at the end of helices were used to obtain MFE values for the 42 genomes.

3.3.4 Calculation of proportion of Shine-Dalgarno in each strain

Gene expression levels were measured by protein abundance based on Pax-DB data and by the I_{TE} index. Of the 26 species in Pax-DB, only 12 species have I_{TE} files in DAMBE. For the remaining 14 bacterial species and all 16 archaeal species, I_{TE} files were created using DAMBE. I_{TE} files for these species were created by selecting 40 ribosomal protein genes, selected as highly expressed genes (HEGs) based on the codon bias database (CBDB) (Hilterbrand *et al.*, 2012). 40 genes with the lowest non-zero protein abundance values selected as lowly expressed genes (LEGs). Codon frequencies of the 40 genes were calculated by DAMBE's relative synonymous codon usage function (Xia, 2013). The frequencies of codon usage of the 40 HEGs and LEGs were then used to create I_{TE} files.

Once I_{TE} values were obtained, the 5' untranslated region (5' UTR) for each genome was also analyzed in DAMBE, using the corrected anti-SD sequence (Table 3.1-3.4). The MFE values, I_{TE} values and the Num match values were sorted from high to low I_{TE} values. Of the total genes within each genome, the top 10% were categorized as (HEGs) and the bottom 10% were

categorized as (LEGs). Within the HEGs and LEGs, counts of 'Num match' of greater than or equal to one were used to define the Shine-Dalgarno containing portion of genes (P_{SD}-HEG, P_{SD}-LEG), where:

$$P_{SD} - \text{HEG} = \frac{\text{Count of 'Num match } \geq 1'}{\text{Total number of HEGs}}$$

$$P_{SD} - \text{LEG} = \frac{\text{Count of 'Num match } \geq 1'}{\text{Total number of LEGs}}$$

Counts of 'Num match' equal to zero were used to define the non-Shine Dalgarno portion of genes (N_{SD}-HEG, N_{SD}-LEG), where

$$N_{SD} - HEG = \frac{\text{Count of 'Num match} = 0'}{\text{Total number of HEGs}}$$

$$N_{SD} - LEG = \frac{\text{Count of 'Num match} = 0'}{\text{Total number of LEGs}}$$

Similarly, average MFE40nt and I_{TE} values for P_{SD}-HEGs, P_{SD}-LEGs, N_{SD}-HEGs and N_{SD}-LEGs and N_{SD}-LEGs were also calculated. The mean number of consecutively matched sites, M_{SD} for each genome was calculated by finding the average length for all SD motif matches (Table 3.5-3.8). Mean base pairing length, distance to AUG and P_{SD}, ITE, and MFE40nt of HEGs and LEGs in all 42 selected species (Gram positive, Gram negative, Euryarchaeota and Crenarchaeota) are listed in Table 3.5-3.8.

3.4 Results and Discussion

Our first objective is to determine whether there is a significant difference in the P_{SD}-HEGs of bacteria and archaea. Our hypothesis is that translation initiation is more efficient for bacterial genes than archaeal genes because SD sequences are more commonly located in the former than

in the latter. The more efficient translation initiation increases codon adaptation and therefore increases the protein production rate in bacteria relative to archaea. Specific predictions are: i) that P_{SD}-HEGs (proportion of highly expressed-SD containing genes) are significantly higher in bacteria than archaea and ii) that M_{SD} (length of SD-aSD pairing) is closer to the optimal length in bacteria than archaea, with the optimal SD length being six nucleotides (Schurr *et al.*, 1993; Komarova *et al.*, 2002; Vimberg *et al.*, 2007), and iii) that MFE40nt upstream of the start codon are less negative in bacteria than in archaea.

3.4.1 Comparison of SD features and secondary structure stability between Bacteria and Archaea

Mean P_{SD}-HEG in 26 bacteria (mean = 0.820) is significantly higher than mean P_{SD}-HEG in 16 archaea (mean = 0.698) ($t = 2.218$, $DF = 32$, $p<0.05$, two-tailed t-test assuming unequal variance) and MFE_{40nt}-HEG in bacteria (mean = -3.62kJ/mol) is also significantly stronger than MFE_{40nt}-HEG of archaea (mean = -5.94kJ/mol) ($t = 4.29$, $DF = 30$, $p<0.001$, two tailed t-test assuming unequal variance). However, there is no significant difference in the P_{SD}-LEGs of bacteria (mean = 0.705) and archaea (mean = 0.668) ($t = 0.682$, $DF=40$, $p>0.05$, t-test assuming equal variance), and MFE_{40nt}-LEG of bacteria (mean = -4.39kJ/mol) and archaea (mean = -6.06kJ/mol) ($t = 2.55$, $DF = 28$, $p>0.01$, two tailed t-test assuming unequal variance).

The mean M_{SD} in both archaea and bacteria is smaller than the optimal six (Schurr *et al.*, 1993; Komarova *et al.*, 2002; Vimberg *et al.*, 2007) but there is no significant difference in M_{SD} of bacteria (mean = 4.78) and archaea (mean = 4.76) ($t = 0.139$, $DF = 40$, $p>0.05$, two tailed t-test assuming equal variance). While P_{SD}-HEG and MFE_{40nt} support our hypothesis that translation initiation is more efficient in bacteria than in archaea but M_{SD} does not.

We then determined if there is a correlation between the two SD features: P_{SD} and M_{SD} and MFE_{40nt} in HEGs, across all the forty two species. There is almost no correlation between: i) P_{SD}-HEG and MFE_{40nt} of the forty two species (r = 0.099) and regression analysis shows no correlation (R^2 = 0.0099, Figure 3.2) ii) MFE_{40nt} and M_{SD} are weakly positively correlated (r = 0.19) and this relation is also almost negligible in the regression analysis (R^2 = 0.0361, Figure 3.3).

To determine whether all or some species show this trend, we further analyzed the correlation between SD features and MFE_{40nt} by selecting the 42 species into four groups: Gram negative, Gram positive, Crenarchaeota and Euryarchaeota.

3.4.2 Analyzing the SD features and secondary structure stability in Gram negative bacteria

There is a significant difference between P_{SD}-HEG (mean= 0.78) and P_{SD}-LEG (mean = 0.64) in Gram-negative bacteria (t-test assuming equal variance: t =2.65 DF= 36, $p<0.05$, two tailed test). We used the t-test with equal variances because the variances were not significantly different from each other according to an F-test (F=1.245, $DF_{numerator}$=18, $DF_{denominator}$ = 18, $p>0.05$). Mean MFE_{40nt}-HEGs (-3.577kJ/mol) is also significantly weaker than mean MFE_{40nt}-LEGs (-4.585kJ/mol) (t=1.976, DF = 34, $p<0.05$, one tailed t-test assuming unequal variance).

In Gram-negative bacteria, there is a very strong correlation between P_{SD}-HEG and MFE_{40nt} (r = 0.977) however, regression analysis show a very weak correlation between the two (R^2 = 0.2846, Figure 3.4). There is a weak negative correlation between MFE_{40nt} and M_{SD} (r = -0.30), but on carrying out a regression analysis (R^2 = 0.0881, Figure 3.5) this weak negative correlation is almost negligible.

Although Gram-negative HEGs have a significantly better codon adaptation, more SD sequences, and weaker secondary structures than LEGs, there is no clear positive relationship

between P_{SD} and MFE_{40nt} and M_{SD} in HEGs. This suggests that selection pressure to maximize translation initiation efficiency occurs across the genome, depending on gene expression level, however the effect of the selection pressure on translation initiation efficiency appears to be either very low or negligible once the genes reach a certain threshold of expression. In addition, factors other than SD sequences may be contributing to translation initiation efficiency.

3.4.3 Analyzing the SD features and secondary structure stability in Gram-positive bacteria

There is no significant difference between P_{SD}-HEG (mean = 0.928) and P_{SD}-LEG (mean = 0.876) in Gram-positive bacteria (t = 0.955, DF = 12, p>0.05, two-tailed t-test assuming equal variance). Also, there is no significant difference between MFE_{40nt}-HEG (mean = -6.05kJ/mol) and MFE_{40nt}-LEG (mean = -5.95kJ/mol) (t = -0.125, DF = 16, p>0.05, two-tailed t-test assuming unequal variance). There is also a very strong correlation between: P_{SD}-HEG and MFE_{40nt} (r = 0.977) which is supported by regression analysis (R^2 = 0.9548, Figure 3.6) and also between MFE_{40nt} and M_{SD} (r = 0.81), which is also supported by regression analysis (R^2 = 0.654, Figure 3.7). As genes becomes highly expressed, mRNA secondary structures becomes weaker.

Although the codon adaptation in HEGs is significantly greater than in LEGs, the proportion of SD sequences (0.928 in P_{SD}-HEG and 0.876 in P_{SD}-LEG) and secondary structure features surrounding the start codon are not significantly different between HEGs and LEGs. In Gram-positive bacteria, there appears to be a positive relationship between the SD features examined: P_{SD}, M_{SD} and MFE_{40nt} such that as P_{SD}-HEG and M_{SD} increase, with P_{SD}-HEG approaching one and M_{SD} closing in on 6 nt, MFE_{40nt}-HEG becomes weaker. This relationship appears to hold true for both HEGs and LEGs of Gram-positive bacteria. Hence this suggests that

in Gram-positive bacteria, selective pressure acts to maximise translation initiation efficiency in both HEGs and LEGs.

3.4.4 Comparison of SD features and secondary structure stability between Gram-positive and Gram-negative bacteria

There is no significant difference in: i) P_{SD}-HEGs of Gram-negative (mean = 0.78) and Gram-positive bacteria (mean = 0.927) (t= -2.09, DF =24, $p<0.05$, two-tailed assuming equal variances), ii) MFE_{40nt}-HEG of Gram-negative (mean = -3.58kJ/mol) and Gram-positive bacteria (mean = -3.73kJ/mol) (t = 0.172, DF = 8, $p>0.05$, two tailed t-test assuming unequal variance), and iii) M_{SD} of Gram-negative (mean = 4.614) and Gram-positive bacteria (mean = 5.216) (t = -4.74, DF = 8, $p>0.001$, two tailed t-test assuming unequal variance).

There is also no significant difference in P_{SD}-LEGs of Gram-negative (mean = 0.641) and Gram-positive (mean = 0.876) (t = -3.67, DF = 24, $p >0.01$, two tailed t-test assuming equal variance) however, there is a significant difference in MFE_{40nt}-LEGs of Gram-negative (mean = -4.59kJ/mol) and Gram-positive bacteria (mean = -3.85kJ/mol) (t = -0.795, DF = 9, $p>0.05$, two tailed t-test assuming unequal variance), with Gram-negative bacteria having a significantly weaker secondary structure than Gram-positive bacteria. These results are in concordance with the previous results in which P_{SD} and MFE_{40nt} of Gram-positive bacteria show no significant difference between HEGs and LEGs.

Comparison of P_{SD}, M_{SD} and MFE_{40nt} refute our hypothesis that translation initiation is more efficient in Gram-positive than in Gram-negative bacteria. Hence translation initiation in Gram-negative bacteria is as efficient as Gram-positive bacteria. This means that SD sequence, with the aid of S1 protein, initiates translation as efficiently as SD sequence alone lacking S1 protein.

3.4.5 Analyzing the SD features and secondary structure stability in *Crenarchaeota*

There is no significant difference in P_{SD}-HEG (mean = 0.621) and P_{SD}-LEG (mean = 0.596) in Crenarchaeota (t = 0.335, DF = 16, p>0.05, two tailed t-test assuming equal variance) and there is also no significant difference in MFE_{40nt}-HEG (mean = -6.05kJ/mol) and MFE_{40nt}-LEG (mean = -5.94kJ/mol) (t = -0.125, DF = 16, p>0.05, two tailed t-test assuming unequal variance).

There is a weak correlation between: P_{SD}-HEG and MFE_{40nt} (r = -0.35) which is even weaker in regression analysis (R^2 = 0.1252, Figure 3.8) and between MFE_{40nt} and M_{SD} (r = 0.362), which is also supported by regression analysis (R^2 = 0.131, Figure 3.9). In Crenarchaeota, while there is almost no relationship between P_{SD}-HEG and MFE_{40nt}, there is also almost no correlation between MFE_{40nt} and M_{SD} such that as M_{SD} increases, MFE_{40nt} becomes weaker (Figure 3.9).

Although codon adaptation in HEGs is better than LEGs, approximately 60% of HEGs and LEGs have SD sequences (mean HEGs = 0.621, mean LEGs = 0.596). Also there is no correlation between P_{SD}-HEG and MFE_{40nt}, and M_{SD}. Therefore, while selection pressure does act to improve codon adaptation, translation initiation by SD mechanism is inefficient in Crenarchaeota.

3.4.6 Analyzing the SD features and secondary structure stability in Euryarchaeota

There is no significant difference in P_{SD}-HEG (mean = 0.798) and P_{SD}-LEG (mean = 0.760) in Euryarchaeota (t = 0.531, DF = 12, p>0.05, two-tailed t-test assuming equal variance) and there is also no significant difference in MFE_{40nt}-HEG (mean = -5.79kJ/mol) and MFE_{40nt}-LEG (mean = -6.21kJ/mol) (t = 0.331, DF = 11, p>0.05, assuming unequal variance).

P_{SD}-HEG and MFE_{40nt} are weakly correlated (r = 0.23) and this relation is almost negligible in regression analysis (R^2 = 0.0528, Figure 3.10). However, there is a fairly weak positive correlation between MFE_{40nt} and M_{SD} (r = 0.646), which is also supported by regression analysis

(R^2 = 0.4168, Figure 3.11). In Euryarchaeota, while there appears to be almost no relationship between P_{SD}-HEG and MFE_{40nt}, but there is a strong negative correlation between MFE_{40nt} and M_{SD} such that as M_{SD} increases, MFE becomes weaker.

Although approximately 76% of the genes have SD sequences in HEGS (mean P_{SD} = 0.798) and LEGs (mean P_{SD} = 0.76) and codon adaptation is better in HEGs, selective pressure appears to weakly act to improve SD translation initiation efficiency in Euryarchaeota.

3.4.7 Comparison of Crenarchaeota and Euryarchaeota

No significant difference is detected between the P_{SD}-HEGs of Crenarchaeota (mean = 0.621) and Euryarchaeota (mean = 0.798) ($t = -2.32$, $DF = 14$, $p<0.05$, two tailed assuming equal variance), but MFE_{40nt}-HEG of Crenarchaeota (mean = -6.05kJ/mol) indicates secondary structure is significantly stronger than for Euryarchaeota (mean = -5.79kJ/mol) ($t = -0.28$, $DF = 12$, $p>0.05$, two tailed t-test assuming unequal variance).

For the lowly expressed genes, there is no significant difference in P_{SD}-LEG of Crenarchaeota (mean = 0.596) and Euryarchaeota (mean = 0.760) ($t = -2.28$, $DF = 14$, $p<0.05$, two-tailed t-test assuming equal variance), and there is also no significant difference in the MFE_{40nt}-LEG of the former (mean = -5.95kJ/mol) and the latter (mean = -6.21kJ/mol) ($t = 0.218$, $DF = 10$, $p>0.05$). There is also no significant difference between M_{SD} of Crenarchaeota (mean = 4.60) and Euryarchaeota (mean = 4.97) ($t = -3.21$, $DF = 14$, $p>0.001$, two tailed t-test assuming equal variance).

3.4.8 Construction of phylogenetic tree of 42 selected species with 16S rRNA genes

To assess whether closely related species tend to have similar SD features, 16S rRNA phylogenetic tree was constructed for all the selected species. 16S ribosomal RNA genes of 42

selected species were extracted by DAMBE. Their 16S ribosomal RNA genes were aligned by Clustalw installed in DAMBE. Phylogenetic tree was constructed by DAMBE using distance based method neighbor-joining with a bootstrapping resampling threshold of 1000 (Figure 3.1).

Numerous studies in translation efficiency show the dependence of codon usage and translation elongation efficiency on translation initiation efficiency (Supek & Smuc, 2010; Tuller *et al.*, 2010; Prabhakaran *et al.*, 2015; Xia, 2015). When translation initiation is inefficient, mRNA conversion to protein is not increased even if codon usage is optimal. However, if translation initiation is efficient, then protein production depends on tRNA availability and mRNA with optimized codon usage can increase protein production. Therefore, genes that have a low translation initiation efficiency undergo less selection pressure than genes having strong translation initiation efficiency.

We used this hypothesis and the knowledge that bacterial genomes tend to have more SD presence than archaeal genomes (Ma *et al.*, 2002) to determine whether translation initiation is more efficient in the former than in the latter. Our hypothesis that bacteria genomes have more efficient translation initiation than archaeal genomes is fairly consistent with our interpretation of the data; in general, bacterial genomes exhibit higher P_{SD} but not M_{SD} than archaeal genomes. Although there is no strong relationship between P_{SD}, $M_{SD,}$ and MFE_{40nt} between bacterial and archaeal genomes, there is a very strong negative correlation between P_{SD} and MFE_{40nt} in Gram-positive bacteria such that, as P_{SD} increases, the MFE_{40nt} nt becomes stronger. The results indicate that, in Gram-positive bacteria, SD-led translation initiation is common in both HEGs and LEGs and is highly efficient in HEGs. The high P_{SD} proportion is in accordance with previous studies which have also found significantly higher P_{SD} (percent of SD in total genes) in Gram-positive than Gram-negative bacteria.

The high P_{SD} in HEGs and LEGs may also explain why the removal of SD sequence in Gram-positive bacteria leads to no protein production. As there is no difference between P_{SD} in HEGs and LEGs, and because we expect LEGs to have the least translation initiation optimization when comparing HEGs, average genes and LEGs, it is safe to assume that the average genes also have high P_{SD} values. Therefore, SD sequences are present in all genes. Also because there is no difference in MFE_{40nt} of HEGs and LEGs, it also means that the secondary structures flanking are similar. Selection pressure acts on the genome, based on gene expression level. However since P_{SD} and MFE_{40nt} are not significantly different in HEGs and LEGs, it appears that the effects of selection pressure are either too low, or are reaching a threshold high. In summary, Gram-positive high levels of gene expression is brought about by an efficient translation initiation by optimal SD sequences.

For the remaining three groups: Gram-negative bacteria, Crenarchaeota and Euryarchaeota, there is little or no relationship between P_{SD}, $M_{SD,}$ and MFE_{40nt} in HEGs. Also, only in Gram-negative bacteria is there a significant difference between the P_{SD} and MFE_{40nt} in HEGs and LEGs, with MFE_{40nt} being significantly weaker in HEGs than LEGs. The results indicate that in Gram-negative bacteria as genes become highly expressed, the selection pressure favors SD presence and weaker secondary structures flanking the start codon.

In *E.coli* and other Gram-negative bacteria, the S1 protein docks and unfolds structured mRNA and also correctly positions the initiation codon inside the decoding channel and is essential for translation initiation in mRNA having weak or degenerate SD sequences (Komarova *et al.*, 2002; Duval *et al.*, 2013). Our results suggest that in Gram-negative bacteria, because SD sequences are required for translation initiation, selection pressure favors both SD presence and weaker secondary structures flanking start codon in HEGs. It is unclear whether the sequences

themselves are optimized in which case SD dependency on S1 protein is reduced, or whether HEGs tend to have weak SD sequences and S1 protein both of which interact to initiate translation. In summary, in Gram-negative bacteria, high levels of gene expression is brought about by the avoidance of a secondary structure at the 5' end of mRNA as well as either optimization of SD sequences or increased presence of weak SD and S1 protein.

The lack of strong correlation between P_{SD}, $M_{SD,}$ and MFE_{40nt} in HEGs and also the lack of significant difference between P_{SD}-HEGs and LEGs, and MFE_{40nt}-HEGs and LEGs, in Crenarchaeota and Euryarchaeota show that even in highly expressed genes, SD-led translation initiation in archaea is inefficient. The lack of significant difference in P_{SD}-HEGs vs P_{SD}-LEGs in both archaea groups also indicates that in these genomes, the selection for SD sequences to reach their optimal is very slow. A slow selection could mean that either SD is not very important for translation initiation hence SD-anti-SD binding could be bypassed during translation initiation or, that it is present and co-ordinates with other factors promoting translation initiation when they are present otherwise not. Our results are in concordance with previous studies which have found that in archaeal genomes leaderless translation initiation as wide spread as SD-led translation initiation (Tolstrup *et al.*, 2000; Kramer *et al.*, 2014). In Euryarchaeota and Crenarchaeota, high gene expression levels appears to be brought about by SD-led mechanism and weaker secondary structure as well as non-SD led mechanism; with the latter mechanism more predominant in Crenarchaeota then Euryarchaeota. These results are similar to a previous study in which the SD-motif was more pronounce in Euryarchaeota than Crenarchaeota (Ma *et al.*, 2002; Karlin *et al.*, 2005). The lower P_{SD}-HEG in Crenarchaeota may be explained by the hypothesis that species exposed to rapid growth have more rRNA operons, more tRNA genes and more strongly selected codon usage bias (Sharp *et al.*, 2005). Upon comparison of the minimum generation of

Crenarchaeota to other groups (Table 3.9), we see Crenarchaeota has more species with large growth times than the other groups hence it has lower selected codon usage bias and hence lower P_{SD}.

Of the seven Gram-positive species, only six of them have P_{SD} greater than 0.90 while only one species*: M.tuberculosis,* has the lowest P_{SD}-HEG (=0.698) and MFE_{40nt} (=-8.67kJ/mol). These values are in concordance with a recent study which used a combination of high-throughput transcriptome and ribosome profiling approaches to understand protein expression in two mycobacterial species: *M.smegmatis* and *M.tuberculosis* found that nearly 25% (of total genome) of the mycobacterial transcripts are leaderless, lacking the 5' UTR and the SD ribosome-binding site, and that the leaderless feature is relatively robust when comparing to leadered initiation (Shell *et al.,* 2015).

To determine whether the regression analysis of the Gram-positive bacteria showed a very strong negative correlation due to an outlier effect (the P_{SD} and MFE_{40nt} values of *M. tuberculosis* skews the relationship between P_{SD}-HEG and MFE_{40nt}), we first carried out a correlation analysis ($r = 0.977$) and then the regression analysis. Since there is no big difference in the r-value from the correlation analysis and the R^2 value of the regression analysis, we have confirmed that the strong negative relative seen is due to a strong correlation and not because of an outlier effect. From the regression analysis, it is apparent that as P_{SD}-HEG increases in Gram-positive genomes, the MFE_{40nt} tends to reach ~-2.8 to -3.0 kJ/mol. To determine why there is such a huge difference between the P_{SD}-HEGs of the six Gram-positive bacteria and *M.tuberculosis,* we compared the lifestyle features of these species (Table 3.9) and noticed that the main difference between them is that *M.tuberculosis* is an Actinobacteria which is an obligate aerobe whose minimum generation time is 19h whereas the other species are facultative bacilli with minimum generation times

between 0.4 and 1h. Hence either all or one of these factors may be contributing to the low P_{SD} content in comparison to the other Gram-positive bacteria. Comparatively, these results also confirm that rapidly synthesizing bacteria undergo stronger codon bias in highly expressed genes that slowly synthesizing bacteria. Our results are consistent with previous studies which found that facultative (with or without oxygen) organisms exhibit the largest codon bias and anaerobic organisms show the smallest values when CAI was used as a gene expression proxy. Bacteria living in multiple environments, outside and within hosts, with and without oxygen tend to have higher codon bias than bacteria living in constant environments. Correlation between growth rate and habitat type was found to be highly dependent on CAI values (Botzman *et al.*, 2011). This explains why the association we see between the high growth rate of *M.tuberculosis* (19h) and the high I_{TE} values for HEGs.

3.5 Conclusions

We wish to highlight our findings in summary that there is a significant difference between the P_{SD}-HEG in bacterial and archaeal genomes. In gram positive bacteria when there is a weak secondary structure, SD sequences becomes stronger and more frequent. In gram negative bacteria whenever secondary structure becomes weak, we see a slight decrease in amount and strength of SD sequence. Further, both types of archaea may use the leaderless mechanisms as well as SD mediated translation initiation. In summation, in Gram-positive bacteria, high gene expression is brought about by efficient translation initiation via optimal SD sequences. In Gram-negative bacteria, high gene expression appears to be brought about by weaker secondary structures at the 5' end of mRNA and efficient translation either through weak SD as well as S1 protein. In Euryarchaeota and Crenarchaeota, high gene expression level appears to be brought about by SD

presence, weak secondary structure at the 5’ end of mRNA, as well as non-SD led translation initiation.

Table 3.1. Details of Gram Negative bacteria selected from PaxDb

Species Name	Phylum	Accession	Anti-SD
Pseudomonas aeruginosa PAO1	Gammaproteobacteria	NC_002516	GATCACCTCCTTA
Bacteroides thetaiotaomicron VPI-5482	Bacteroidetes	NC_004663	GAACACCTCCTTT
Deinococcus deserti VCD115	Deinococcus-Thermus	CP001114	GATCACCTCCTTT
Leptospira interrogans serovar Lai str. 56601	Spirochaetes; Leptospirales	NC_005823	GATCACCTCCTTT
Legionella pneumophila sub str. Philadelphia	Gammaproteobacteria	NC_002942	GATCACCTCCTTA
Salmonella enterica serovar Typhimurium	Gammaproteobacteria	NC_003197	GATCACCTCCTTA
Escherichia coli str. K-12 substr. MG1655	Gammaproteobacteria	NC_000913	GATCACCTCCTTA
Helicobacter pylori_26695	Epsilonproteobacteria	NC_000915	GATCACCTCCTTA
Mycoplasma pneumoniae_M129	Tenericutes; Mollicutes	NC_017504.1	GATCACCTCCTTT
Neisseria meningitidis MC58	Betaproteobacteria	NC_003112	GATCACCTCCTTT
Synechocystis sp. PCC 6803	Cyanobacteria	NC_017277	GATCACCTCCTTT
Yersinia pestis CO92	Gammaproteobacteria	NC_003143	GATCACCTCCTTA
Desulfovibrio vulgaris str. Hildenborough	Deltaproteobacteria	NC_002937	GATCACCTCCTTT
Microcystis aeruginosa NIES-843	Cyanobacteria	AP009552	GATCACCTCCTTA
Shewanella oneidensis MR-1	Gammaproteobacteria;	NC_004347	GATCACCTCCTTA
Shigella flexneri 2a str. 301	Gammaproteobacteria	NC_004337	GATCACCTCCTTA
Bartonella henselae strain Houston-1	Alphaproteobacteria	NC_005956	GATCACCTCCTTT
Campylobacter jejuni subsp. NCTC 11168	Epsilonproteobacteria	NC_002163	GATCACCTCCTTA

(1)Anti-SD-Anti shine Dalgarno sequence

Table 3.2. Details of Gram Positive Bacteria selected from PaxDb

Species Name	Phylum	Accession	Anti-SD
Lactococcus lactis subsp. lactis Il1403	Firmicutes; Bacilli	NC_002662	GATCACCTCCTTT
Listeria monocytogenes EGD-e	Firmicutes; Bacilli;	NC_003210	GATCACCTCCTTT
Bacillus subtilis subsp. subtilis str. 168	Firmicutes; Bacilli	NC_000964	GATCACCTCCTTT
Staphylococcus aureus mu3	Firmicutes; Bacilli	BA000017	GATCACCTCCTTT
Bacillus anthracis str. Sterne	Firmicutes; Bacilli	NC_005945	GATCACCTCCTTT
Mycobacterium tuberculosis H37Rv	Actinobacteria	NC_000962	GATCACCTCCTTT
Streptococcus pyogenes M1 GAS	Firmicutes; Bacilli	NC_002737	GATCACCTCCTTT

(1)Anti-SD-Anti shine Dalgarno sequence

Table 3.3. Details of Crenarchaeota Species

Species Name	Phylum	Accession	Anti-SD
Aeropyrum pernix K1	Crenarchaeota	NC_000854.2	GAUCACCUCCCGA
Sulfolobus solfataricus P2	Crenarchaeota	NC_002754.1	GAUCACCUCAUAG
Sulfolobus tokodaii str. 7	Crenarchaeota	NC_003106.2	GAUCACCUCACAU
Metallosphaera sedula DSM 5348	Crenarchaeota	NC_009440.1	GAUCACCUCACAU
Pyrobaculum aerophilum str. IM2	Crenarchaeota	NC_003364.1	GAUCACCUCCACC
Pyrobaculum arsenaticum DSM 13514	Crenarchaeota	NC_009376.1	GAUCACCUCCACC
Thermofilum pendens Hrk 5	Crenarchaeota	NC_008698.1	GAUCACCUCCUUU
Pyrobaculum islandicum DSM 4184	Crenarchaeota	NC_008701.1	GAUCACCUCCACC
Pyrobaculum neutrophilum V24Sta	Crenarchaeota	NC_010525.1	GAUCACCUCCACC

(1)Anti-SD-Anti shine Dalgarno sequence

Table 3.4. Details of Euryarcheota Species

Species Name	Phylum	Accession	Anti-SD
Archaeoglobus fulgidus DSM 4304	Euryarchaeota	NC_000917.1	GAUCACCUCCUAA
Halobacterium sp. NRC-1	Euryarchaeota	AE004437	GATCACCTCCTAA
Methanocaldococcus jannaschii DSM 2661	Euryarchaeota	NC_000909.1	GAUCACCUCCUAA
Methanopyrus kandleri AV19	Euryarchaeota	NC_003551.1	GAUCACCUCCAGC
Pyrococcus furiosus DSM 3638	Euryarchaeota	NC_003413.1	GAUCACCUCCUAU
Pyrococcus horikoshii OT3	Euryarchaeota	NC_000961.1	GAUCACCUCCUAU
Thermococcus gammatolerans EJ3	Euryarchaeota	CP001398	GATCACCTCCTAT

(1)Anti-SD-Anti shine Dalgarno sequence

Table 3.5. Mean base pairing length, distance to AUG and P_{SD}, I_{TE}, and MFE40nt of HEGs and LEGs in Gram Negative Bacteria

Species(Paxdb)	M_{SD}	D_{toAUG}	P_{SD}	I_{TE}	MFE_{40nt} (kJ/mol)	P_{SD}	I_{TE}	MFE_{40nt} (kJ/mol)
				HEG			**LEG**	
Acidithiobacillus ferrooxidans	4.767	14	0.885	0.815	-5.039	0.659	0.747	-6.903
Pseudomonas aeruginosa PAO1	4.63	14	0.806	0.419	-5.337	0.836	0.28	-7.603
Bacteroides thetaiotaomicron 5482	4.34	12	0.376	0.688	-0.18	0.43	0.458	-2.555
Deinococcus deserti VCD115	4.817	15	0.878	0.79	-5.892	0.519	0.549	-8.307
Leptospira interrogans serovar Lai	4.516	14	0.851	0.79	-3.141	0.474	0.672	-3.389
LegionellapneumophilaPhiladelphia	4.567	15	0.721	0.823	-3.259	0.721	0.74	-3.779
Salmonella enterica. serovar	4.654	13	0.863	0.549	-3.79	0.737	0.317	-4.753
Escherichia coli str. K-12MG1655	4.689	13	0.884	0.651	-3.752	0.775	0.416	-4.225

Helicobacter pylori 26695	4.844	12	0.882	0.843	-2.387	0.818	0.777	-3.327
Mycoplasma pneumoniae M129	4.607	17	0.488	0.885	-1.746	0.317	0.812	-3.885
Neisseria meningitidis MC58	4.691	12	0.959	0.622	-3.154	0.631	0.382	-3.742
Synechocystis sp. PCC 6803	4.425	15	0.552	0.743	-3.388	0.506	0.636	-3.685
Yersinia pestis CO92	4.599	13	0.879	0.614	-3.597	0.733	0.4	-4.43
Desulfovibrio vulgaris	4.85	13	0.952	0.661	-5.789	0.757	0.473	-7.396
Microcystis aeruginosa NIES-843	4	15	0.47	0.696	-3.814	0.405	0.488	-4.296
Shewanella oneidensis MR-1	4.578	13	0.818	0.605	-3.786	0.786	0.386	-4.213
Shigella flexneri 2a str. 301	4.692	13	0.862	0.502	-3.997	0.582	0.205	-5.18
Bartonella henselae str Houston-1	4.572	13	0.775	0.764	-3.882	0.725	0.631	-3.029
Campylobacter jejuni NCTC 11168	4.838	12	0.913	0.807	-2.036	0.783	0.655	-2.422

(1) M_{SD} - Mean number of consecutively matched sites. (2) D_{toAUG}- Distance from SD to start codon , (3) P_{SD}-Proportion of SD-containing genes, (4) I_{TE} Index of Translation Elongation, (5) *MFE*- minimum folding energy

Table 3.6. Mean base pairing length, distance to AUG and P_{SD}, I_{TE}, and MFE40nt of HEGs and LEGs in Gram positive Bacteria

Species(Paxdb)	M_{SD}	D_{toAUG}	P_{SD}	I_{TE}	MFE_{40nt} (kJ/mol)	P_{SD}	I_{TE}	MFE_{40nt} (kJ/mol)
				HEG			**LEG**	
Lactococcus lactis subsp. Il1403	4.997	13	0.959	0.659	-2.477	0.863	0.293	-2.358
Listeria monocytogenes EGD-e	5.489	15	0.984	0.763	-2.748	0.968	0.541	-2.918
Bacillus subtilis subsp. str. 168	5.457	15	0.976	0.576	-3.095	0.948	0.315	-4.186
Staphylococcus aureus mu3	5.4	16	0.994	0.628	-3.002	0.917	0.346	-2.775
Bacillus anthracis str. Sterne	5.301	15	0.923	0.633	-3.303	0.893	0.337	-2.769
Mycobacterium tuberculosis H37Rv	4.615	14	0.698	0.891	-8.666	0.673	0.809	-8.712
Streptococcus pyogenes M1 GAS	5.256	14	0.961	0.708	-2.82	0.874	0.356	-3.201

(1) M_{SD} - Mean number of consecutively matched sites. (2) D_{toAUG}- Distance from SD to start codon , (3) P_{SD}-Proportion of SD-containing genes, (4) I_{TE} Index of Translation Elongation, (5) *MFE*- minimum folding energy

Table 3.7. Mean base pairing length, distance to AUG and P_{SD}, I_{TE}, and MFE40nt of HEGs and LEGs in Crenarchaeota

Species(Paxdb)	M_{SD}	D_{toAUG}	P_{SD}	I_{TE}	MFE_{40nt} (kJ/mol)	P_{SD}	I_{TE}	MFE_{40nt} (kJ/mol)
				HEG			**LEG**	
Aeropyrum pernix K1	4.637	17	0.959	0.859	-7.877	0.871	0.727	-6.881
Sulfolobus solfataricus P2	4.786	17	0.551	0.765	-3.739	0.65	0.648	-3.658
Sulfolobus tokodaii str. 7	4.67	17	0.587	0.75	-4.11	0.562	0.636	-3.097

Metallosphaera sedula DSM 5348	4.678	16	0.69	0.775	-4.546	0.629	0.674	-4.457
Pyrobaculum aerophilum str. IM2	4.436	12	0.458	0.748	-6.277	0.454	0.648	-6.153
Pyrobaculum arsenaticum 13514	4.514	12	0.504	0.771	-7.251	0.461	0.659	-6.742
Thermofilum pendens Hrk 5	4.607	17	0.785	0.798	-7.531	0.774	0.679	-8.194
Pyrobaculum islandicum 4184	4.433	12	0.505	0.748	-5.375	0.41	0.644	-6.02
Pyrobaculum neutrophilum V24Sta	4.64	15	0.548	0.785	-7.776	0.553	0.662	-8.341

(1) M_{SD} - Mean number of consecutively matched sites. (2) D_{toAUG}- Distance from SD to start codon , (3) P_{SD}-Proportion of SD-containing genes, (4) I_{TE} Index of Translation Elongation, (5) *MFE*- minimum folding energy

Table 3.8. Mean base pairing length, distance to AUG and P_{SD}, I_{TE}, and MFE40nt of HEGs and LEGs in Euryarchaeota

Species(Paxdb)	M_{SD}	D_{toAUG}	P_{SD}	I_{TE}	MFE_{40nt} ***(kJ/mol)***	P_{SD}	I_{TE}	MFE_{40nt} ***(kJ/mol)***
				HEG			**LEG**	
Archaeoglobus fulgidus DSM 4304	5.056	15	0.707	0.8	-4.649	0.703	0.696	-4.678
Halobacterium sp. NRC-1	4.475	17	0.558	0.777	-7.56	0.515	0.661	-10.09
Methanocaldococcus jannaschii DSM 2661	4.882	15	0.754	0.732	-3.397	0.795	0.644	-3.531
Methanopyrus kandleri AV19	4.629	15	0.811	0.786	-9.183	0.769	0.673	-9.885
Pyrococcus furiosus DSM 3638	5.26	16	0.902	0.758	-4.649	0.872	0.652	-4.555
Pyrococcus horikoshii OT3	5.321	16	0.944	0.774	-4.746	0.91	0.661	-4.484
Thermococcus gammatolerans EJ3	5.164	15	0.912	0.805	-6.331	0.759	0.693	-6.246

(1) M_{SD} - Mean number of consecutively matched sites. (2) D_{toAUG}- Distance from SD to start codon , (3) P_{SD}-Proportion of SD-containing genes, (4) I_{TE} Index of Translation Elongation, (5) *MFE*- minimum folding energy

Table 3.9. Characterization of 42 genomes by: (i) Phylum, class and order, (ii) Gram stain, (iii) Temperature conditions at which specie lives, (iv) Sporulation nature, (v) Habitat, (vi) Oxygen requirements, (vii) Optimal temperature (°C)

Strain; Phylum → Class → Order	Family	Gram stain (+/-)	Temp conditions	S/NS	Habitat	Oxygen	Optimal temperature (°C)	Minimum Generation Time(h)	References
Bacteria									
Actinobacteria									
Corynebacteriales									
MYCTUB	Mycobacteriaceae	+	M	NS	HA	OAE	37	19	(Dunn & North, 1995)
Bacteroidetes/Chlorobi									
Bacteriodales									
BACTHE VPI-5482	Bacteridaceae	-	M		HA	OAN	37	1.47	(Anderson & Salyers, 1989)
Cyanobacteria									
Chroococcales									
MICAER - NIES-843	Microcystis	-	M		AQ	OAE			
Synechocytosis- sp. PCC 6803	Synechocystis	-	M	NS	AQ	FA			
Deinococcus – Thermus									
Deinococcales									
DEIDES - VCD115	Deinococcaceae	-	M	NS	TE	OAN			
Firmicutes									
Bacillales									
BACANT - str. Sterne	Bacillaceae	+	M	S	MU	FA	37	.5	(Chakrabarty et al., 2006)
BACSUB - subsp. subtilis str. 168	Bacillaceae	+	M	S	TE	FA	37	0.43	(Hogness et al., 1964)

LISMON - EGD-e	Listeriaceae	+	M	NS	MU	FA	23	1	(Rubin, 1986)
STAAUR - subsp. aureus Mu3	Staphylococcaceae	+	M	NS	HA	FA			
Lactobacillales									
LACLAC - subsp. lactis Il1403	Streptococcaceae	+	M	NS	MU	FA	40	0.7	(Andersen et al., 2001)
STRPYO - M1 GAS	Streptococcaceae	+	M	NS	HA	FA	37	0.4	(Biswas, Germon, McDade, & Scott, 2001)
Proteobacteria **α-proteobacteria** **Rhizobales**									
BARHEN - strain Houston-1	Bartonellaceae	-	M		HA	OAE	37	3	(Chenoweth, Somerville, Krause, Reilly, & Gherardini, 2004)
β – proteobacteria **Neisseriales**									
NEIMEM - MC58	Neisseriaceae	-	M	NS	HA	OAE	35		
Δ proteobacteria **Desulfovibrionales**									
DESVUL - str. Hildenborough	Desulfovibrionaceae	-	M	NS	MU	OAN	32	14	(Pohorelic et al., 2002)
ϒ proteobacteria **Acidithiobacillales**									
ACIFER - ATCC 23270	Acidithiobacillaceae	-	M	NS	TE	FA	30		
Alteromonadales									
SHEONE - MR-1	Shewanellaceae	-	M	NS		FA	30	0.66	(Abboud et al., 2005)
Enterobacteriales									

ESCCOL - str. K-12 substr. MG1655	Enterobacteriaceae	-	M	NS	HA	FA	37	0.35	(Rubin, 1986)
SALENT - subsp. enterica serovar Typhimurium str. LT2	Enterobacteriaceae	-	M	NS	HA	FA	37		
SHIFLE - 2a str. 301	Enterobacteriaceae	-	M	NS	HA	FA	37		
YERPES - CO92	Enterobacteriaceae	-	M	NS	MU	FA	37	1.70	
Legionellales									
LEGPNE subsp. Pneumophila str. Philadephia 1	Legionellaceae	-	M	NS	HA	OAE	37	3.3	
Pseudomondales									
PSEAER PAO1	Pseudomonodaceae	-	M	NS	MU	OAE	37	0.5	(Rubin, 1986)
ε proteobacteria									
Camplyobacterales									
CAMJEJ - subsp. jejuni NCTC 11168 *	Campylobacteraceae	-	M	NS	MU	MI	37	1.5	(Rollins, Coolbaugh, Walker, & Weiss, 1983)
HELPYL – 26695	Heliobacteraceae	-	M	NS	HA	OAE	37	2.4	(Vega, Cortiñas, Mattana, Silva, & Puig De Centorbi, 2003)
Spirochaetes									
Spirochaetia									
Leptospirales									
LEPINT - serovar Lai str. 56601	Leptospiraceae	-	M		HA	OAE	29	9	(Shenberg, 1967)
Tenericutes									
Mollicutes									
Mycoplasmatales									
MYCPNE M129	Mycoplasmataceae	-	M	NS	HA	FA	37	6	(Yus et al., 2009)

Archaea									
Crenarchaeota									
Thermoprotei									
Desulfurococcales									
AERPER - K1	Desulfurococcaceae		H		SP	FA	95	4	(Robinson & Bell, 2007)
Sulfobales									
METSED - DSM 5348	Sulfolobaceae		TA	NS	SP	OAE	70		
SULSOL - P2	Sulfolobaceae						87	6	(Torarinsson, Klenk, & Garrett, 2005)
SULTOK - str. 7	Sulfolobaceae		H	NS	SP	OAE	80	6	(Torarinsson et al., 2005)
Thermoproteales									
THEPEN - Hrk 5	Thermofilaceae		H	NS	SP	OAN	88		
PYRAER - str. IM2	Thermoproteaceae						100	3	(Torarinsson et al., 2005)
PYRARS - DSM 13514	Thermoproteaceae		H	NS	SP	OAN	95	1.3	(Niggemyer, Spring, & Stackebrandt, 2001)
PYRISL - DSM 4184	Thermoproteaceae		T	NS	SP	OAN	100		
PYRNEU - V24Sta									
Thermoproteaceae									
Euryarchaeota									
Archaeoglobi									
Archeoglobales									
ARCFUL - DSM 4304	Archaeoglobaceae		H		AQ	OAN	83		
Halobacteria									
Halobacteriales									
Halobacterium sp. NRC-1	Halobacteriaceae		M	NS	SP	FA	37	9	(Woodson, Peck, Krebs, & Escalante-Semerena, 2003)

Methanococci **Methanococcales**									
METJAN - DSM 2661	Methanocaldococcaceae		H	NS	AQ	OAN	83	0.5	(Torarinsson et al., 2005)
Methanopyri **Methanopyrales**									
METKAN - AV19	Methanopyraceae		H	NS	SP	OAN	98	0.83	(Rospert et al., 1991)
Thermococci **Thermococcales**									
PYRFUR - DSM 3638	Thermococcaceae		H	NS	AQ	OAN	100	0.62	(Torarinsson et al., 2005)
PYRHOR - OT3	Thermococcaceae		H	NS	AQ	OAN	98	0.62	(Torarinsson et al., 2005)
THEGAM - EJ3	Thermococcaceae		H	NS	SP	OAN	88		

(i) Gram stain: + indicates gram positive, - indicates Gram-negative
(ii) Temperature conditions: H stands for hyperthermophilic, M for mesophilic, and T for thermophilic.
(iii) Sporulation: NS stands for non-sporulating and S for sporulating
(iv) Habitat: HA stands for host associated, AQ for aquatic, SP for specialized, TE for terrestrial and MU for multiple
(v) Oxygen requirements: AE stands for aerobic, OAE stands for obligate aerobe, OAN, for obligate anaerobe, FA for facultative and MI for microaerophilic.
(vi) References for minimum generation time and optimal temperature

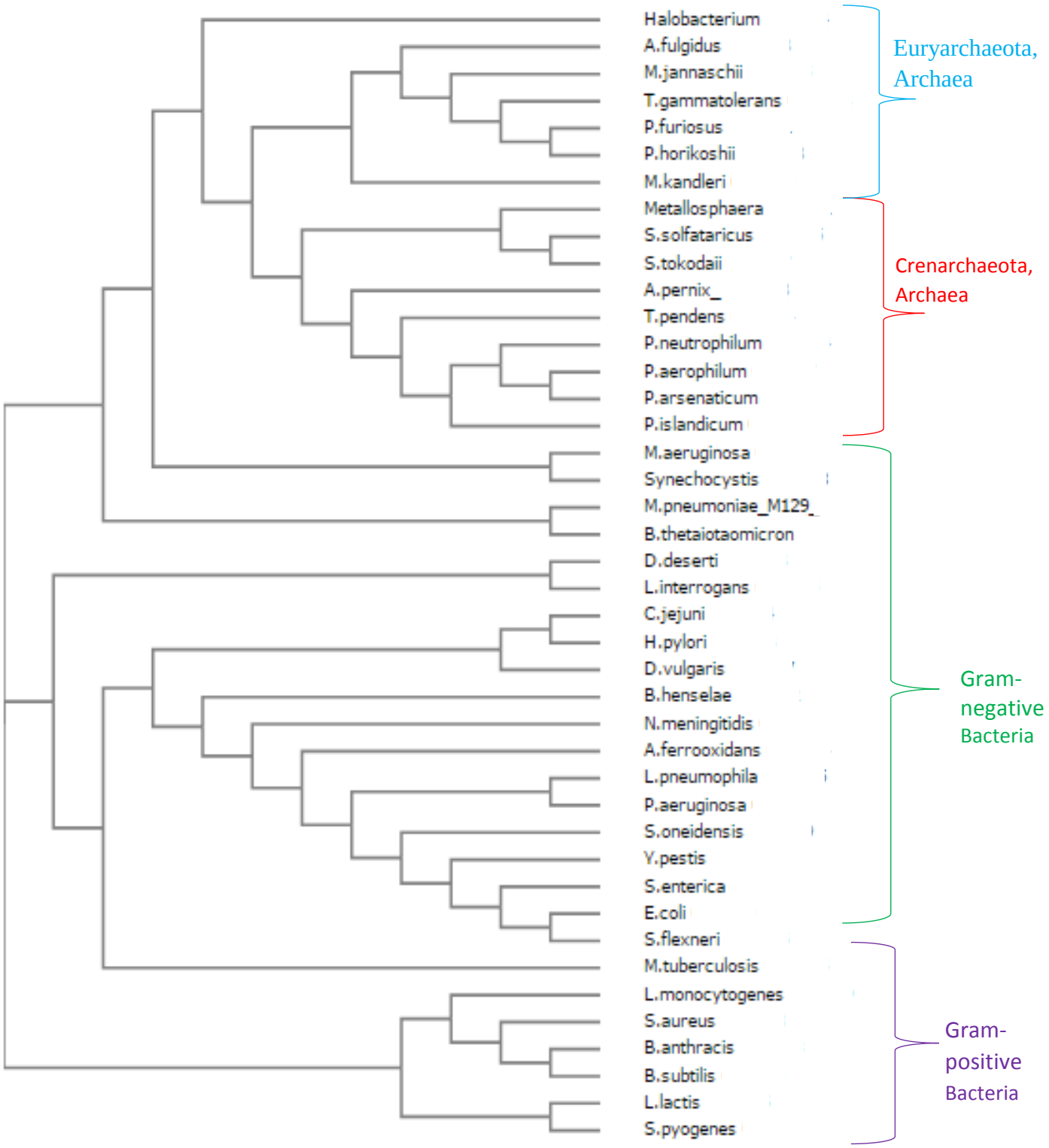

Figure 3.1.Phylogeny tree of 42 genomes based on 16S rRNA sequence created using DAMBE

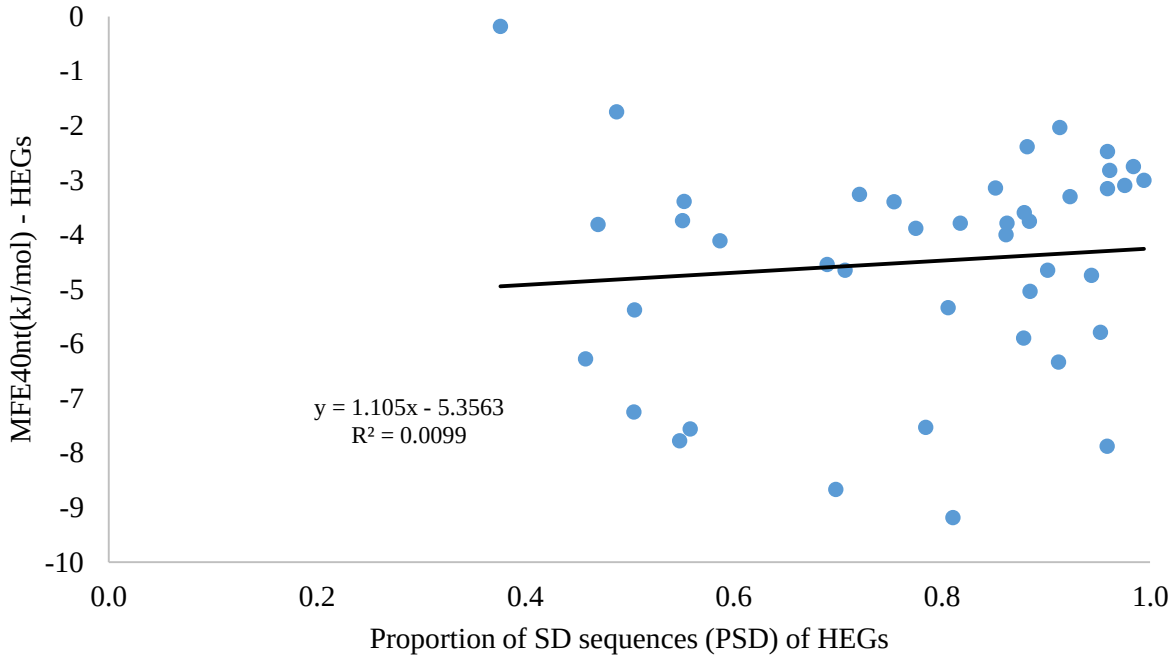

Figure 3.2. Almost no correlation between MFE40nt and P_{SD}- HEGS in forty two species such that as P_{SD} increases, MFE40nt becomes weaker

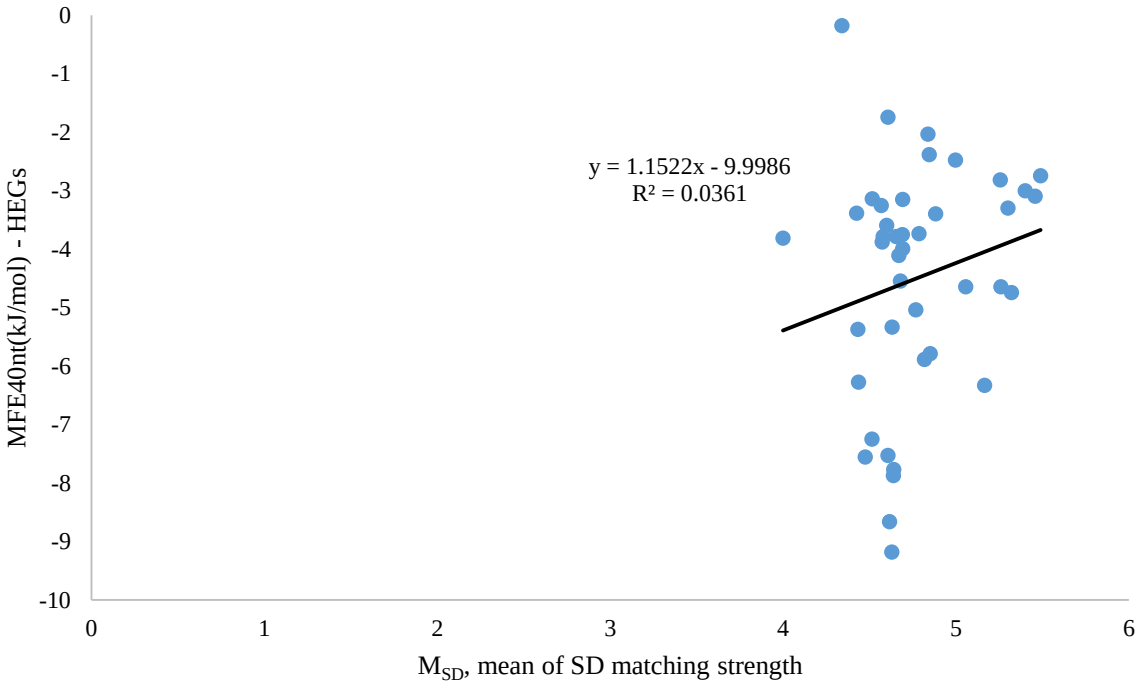

Figure 3.3. No correlation between MFE40nt and M_{SD} for HEGs of all forty two species

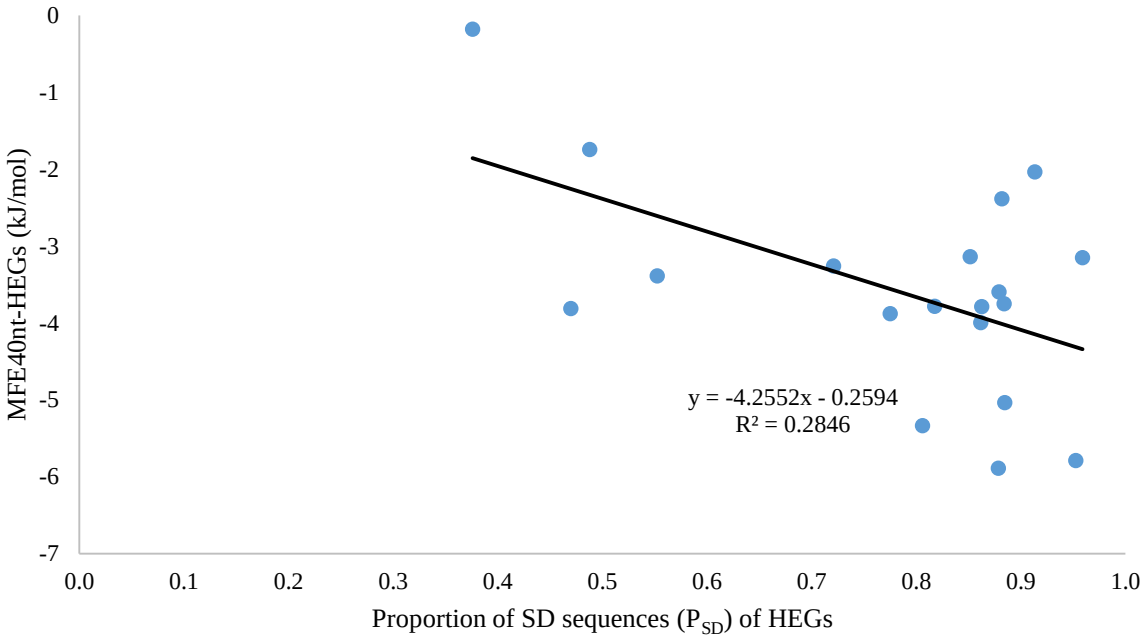

Figure 3.4. Very weak negative correlation between MFE40nt and P_{SD}- HEGS in nineteen gram negative bacterial species such that as P_{SD} increases, RNA secondary structure becomes stronger.

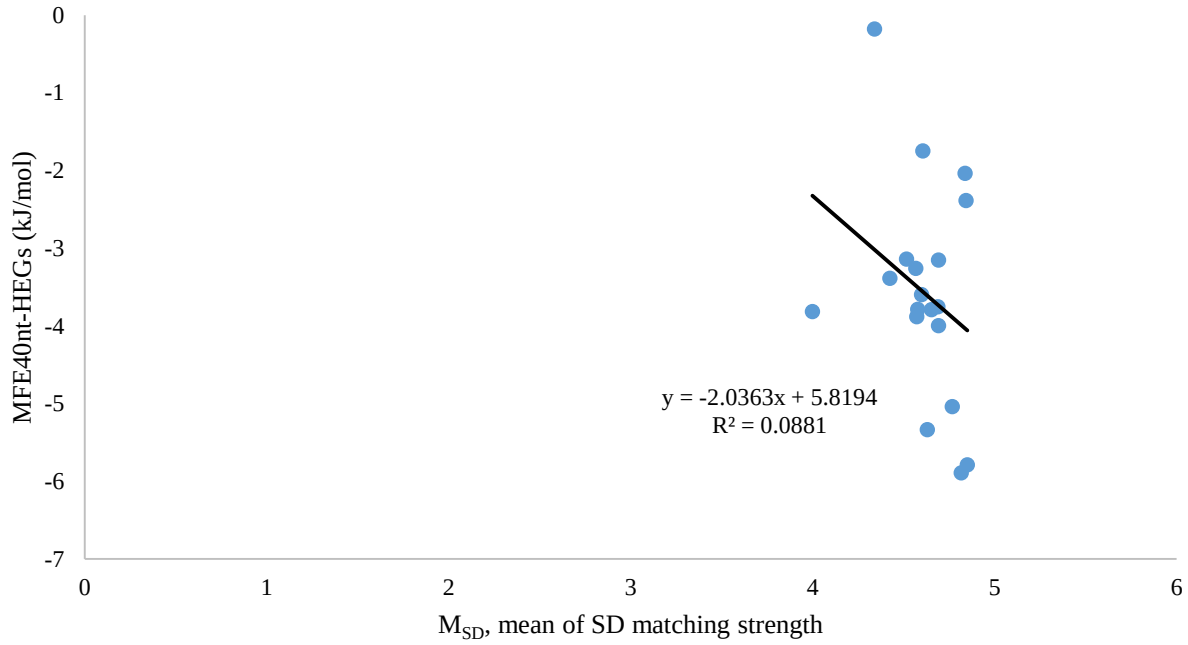

Figure 3.5. Very weak correlation between MFE40nt and M_{SD} for HEGs of all nineteen gram

negative bacterial species such that there is almost no increase in M_{SD} as MFE40nt becomes Stronger

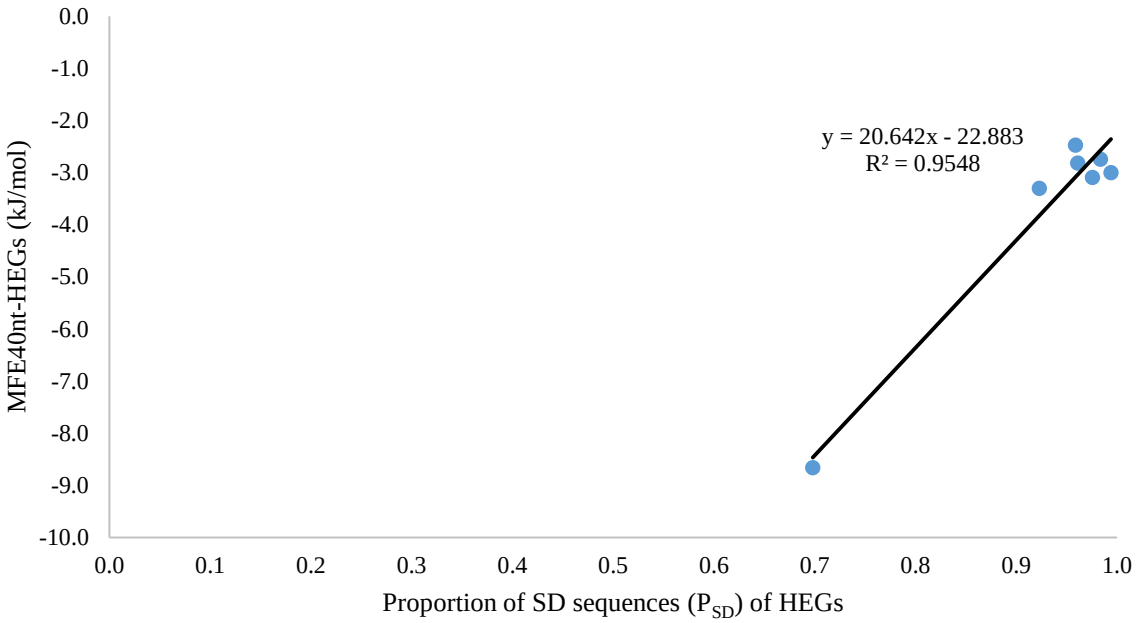

Figure 3.6. Very strong positive correlation between MFE40nt and P_{SD}-HEGs in seven gram positive bacterial species such that as P_{SD} increases, MFE40nt becomes weaker

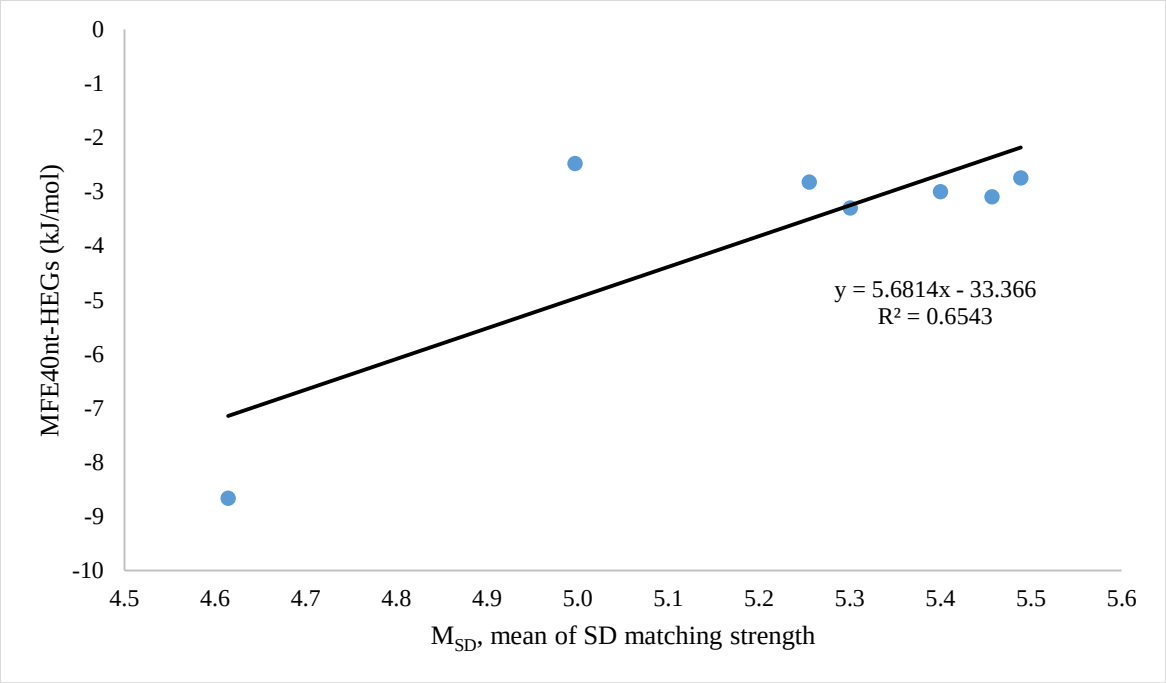

Figure 3.7. A fairly strong positive correlation between MFE40nt and M_{SD} for HEGs of all seven gram positive bacterial species such that as M_{SD} increases, MFE40nt becomes weaker

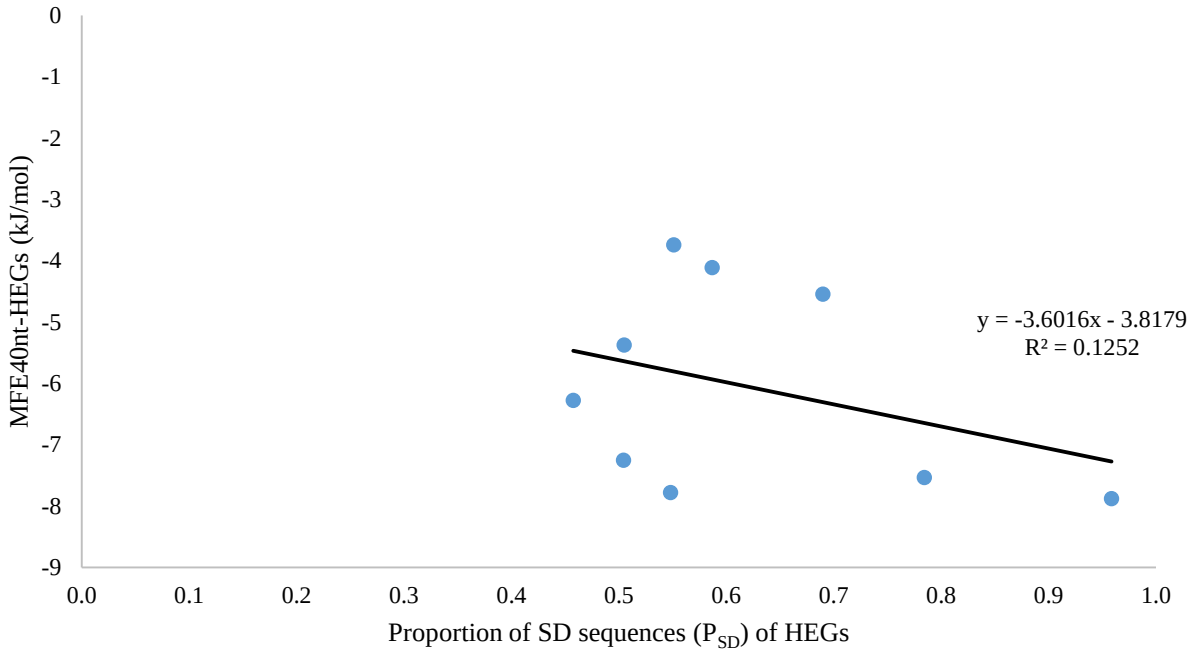

Figure 3.8. Very weak negative correlation between P_{SD}-HEG and MFE40nt-HEG (kJ/mol) of nine Crenarchaeota such that as P_{SD} increases, MFE40nt becomes stronger

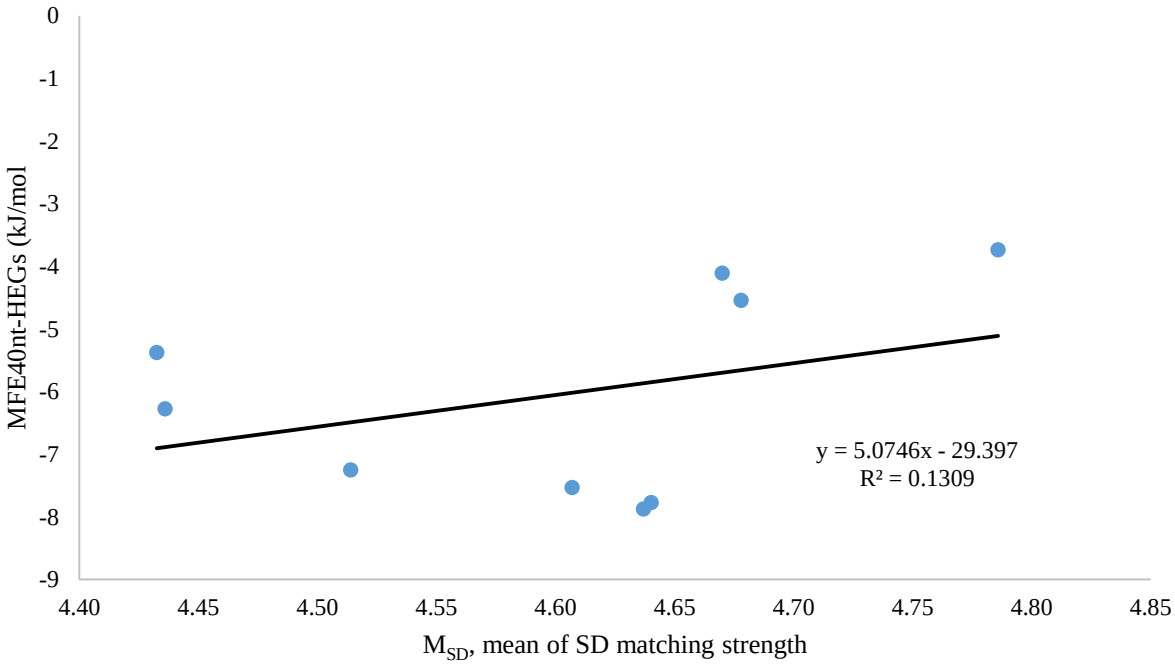

Figure 3.9. Very weak positive correlation between MFE40nt and M_{SD} for HEGs of nine Crenarchaeota species such that as M_{SD} increases, MFE40nt becomes slightly weaker

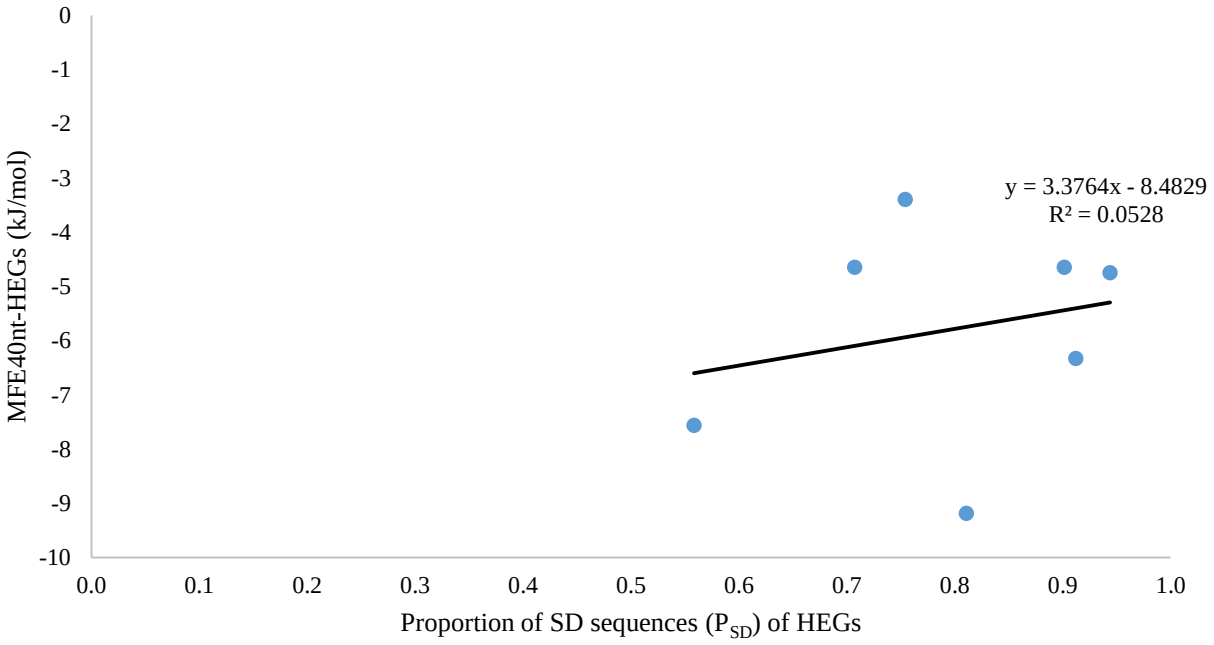

Figure 3.10. Almost no correlation between P_{SD}-HEG and MFE40nt-HEG (kJ/mol) of nine Euryarchaeota

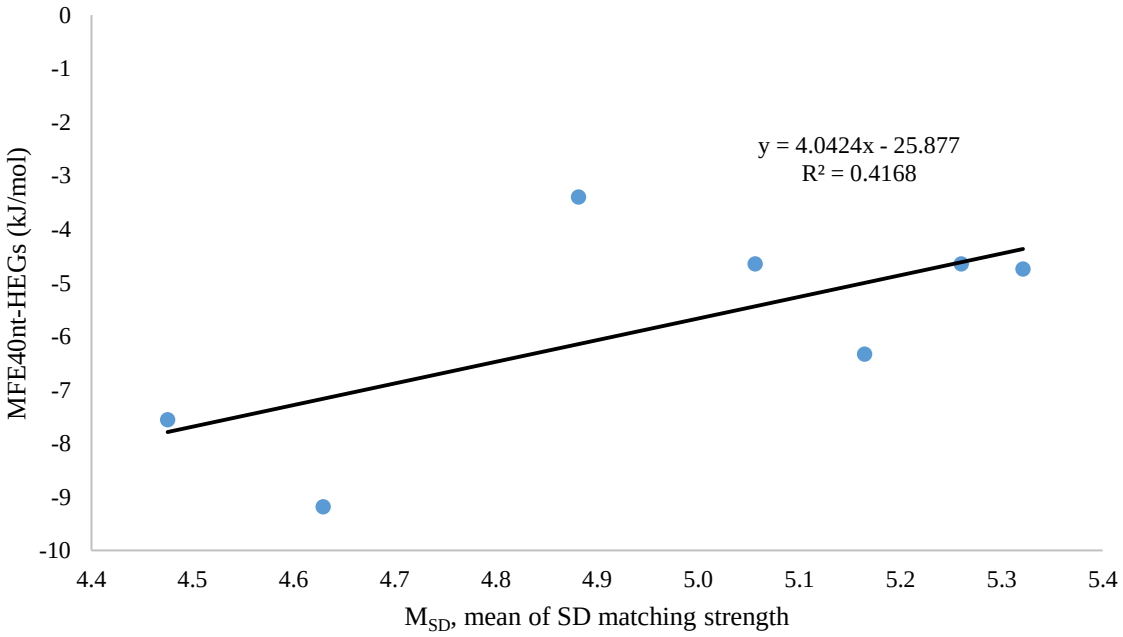

Figure 3.11. Fairly strong positive correlation between MFE40nt and M_{SD} for HEGs of nine Euryarchaeota species such that as M_{SD} increases, MFE40nt becomes weaker

References

Acinas, S. G., Marcelino, L. A., Klepac-Ceraj, V., & Polz, M. F. (2004). Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons. *J Bacteriol, 186*(9), 2629-2635.

Akashi, H. (1994). Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. *Genetics, 136*(3), 927-935.

Band, L., & Henner, D. J. (1984). Bacillus subtilis requires a "stringent" Shine-Dalgarno region for gene expression. *DNA, 3*(1), 17-21.

Barraud, P., Schmitt, E., Mechulam, Y., Dardel, F., & Tisne, C. (2008). A unique conformation of the anticodon stem-loop is associated with the capacity of tRNAfMet to initiate protein synthesis. *Nucleic Acids Res, 36*(15), 4894-4901. doi:10.1093/nar/gkn462

Barrick, D., Villanueba, K., Childs, J., Kalil, R., Schneider, T. D., Lawrence, C. E., . . . Stormo, G. D. (1994). Quantitative analysis of ribosome binding sites in E.coli. *Nucleic Acids Res, 22*(7), 1287-1295.

Benelli, D., & Londei, P. (2011). Translation initiation in Archaea: conserved and domain-specific features. *Biochem Soc Trans, 39*(1), 89-93. doi:10.1042/BST0390089

Bennetzen, J. L., & Hall, B. D. (1982). Codon selection in yeast. *J Biol Chem, 257*(6), 3026-3031.

Bentele, K., Saffert, P., Rauscher, R., Ignatova, Z., & Bluthgen, N. (2013). Efficient translation initiation dictates codon usage at gene start. *Mol Syst Biol, 9*, 675. doi:10.1038/msb.2013.32

Berk, V., Zhang, W., Pai, R. D., & Cate, J. H. (2006). Structural basis for mRNA and tRNA positioning on the ribosome. *Proc Natl Acad Sci U S A, 103*(43), 15830-15834. doi:10.1073/pnas.0607541103

Bottger, E. C. (1989). Rapid determination of bacterial ribosomal RNA sequences by direct sequencing of enzymatically amplified DNA. *FEMS Microbiol Lett, 53*(1-2), 171-176.

Botzman, M., & Margalit, H. (2011). Variation in global codon usage bias among prokaryotic organisms is associated with their lifestyles. *Genome Biol, 12*(10), R109. doi:10.1186/gb-2011-12-10-r109

Brenneis, M., Hering, O., Lange, C., & Soppa, J. (2007). Experimental characterization of Cis-acting elements important for translation and transcription in halophilic archaea. *PLoS Genet, 3*(12), e229. doi:10.1371/journal.pgen.0030229

Britton, R. A., Wen, T., Schaefer, L., Pellegrini, O., Uicker, W. C., Mathy, N., . . . Condon, C. (2007). Maturation of the 5' end of Bacillus subtilis 16S rRNA by the essential ribonuclease YkqC/RNase J1. *Mol Microbiol, 63*(1), 127-138. doi:10.1111/j.1365-2958.2006.05499.x

Brodersen, D. E., Carter, A. P., Clemons, W. M., Jr., Morgan-Warren, R. J., Murphy, F. V. t., Ogle, J. M., . . . Ramakrishnan, V. (2001). Atomic structures of the 30S subunit and its complexes with ligands and antibiotics. *Cold Spring Harb Symp Quant Biol, 66*, 17-32.

Brodersen, D. E., Clemons, W. M., Jr., Carter, A. P., Wimberly, B. T., & Ramakrishnan, V. (2002). Crystal structure of the 30 S ribosomal subunit from Thermus thermophilus: structure of the proteins and their interactions with 16 S RNA. *J Mol Biol, 316*(3), 725-768. doi:10.1006/jmbi.2001.5359

Brosius, J., Palmer, M. L., Kennedy, P. J., & Noller, H. F. (1978). Complete nucleotide sequence of a 16S ribosomal RNA gene from Escherichia coli. *Proc Natl Acad Sci U S A, 75*(10), 4801-4805.

Bulmer, M. (1987). Coevolution of codon usage and transfer RNA abundance. *Nature, 325*(6106), 728-730. doi:10.1038/325728a0

Bulmer, M. (1991). The selection-mutation-drift theory of synonymous codon usage. *Genetics, 129*(3), 897-907.

Calogero, R. A., Pon, C. L., Canonaco, M. A., & Gualerzi, C. O. (1988). Selection of the mRNA translation initiation region by Escherichia coli ribosomes. *Proc Natl Acad Sci U S A, 85*(17), 6427-6431.

Cammarano, P., Mazzei, F., Londei, P., Teichner, A., de Rosa, M., & Gambacorta, A. (1983). Secondary structure features of ribosomal RNA species within intact ribosomal subunits and efficiency of RNA-protein interactions in thermoacidophilic (Caldariella acidophila, Bacillus acidocaldarius) and mesophilic (Escherichia coli) bacteria. *Biochim Biophys Acta, 740*(3), 300-312.

Capecchi, M. R., & Klein, H. A. (1969). Characterization of three proteins involved in polypeptide chain termination. *Cold Spring Harb Symp Quant Biol, 34*, 469-477.

Carter, A. P., Clemons, W. M., Brodersen, D. E., Morgan-Warren, R. J., Wimberly, B. T., & Ramakrishnan, V. (2000). Functional insights from the structure of the 30S ribosomal subunit and its interactions with antibiotics. *Nature, 407*(6802), 340-348. doi:10.1038/35030019

Caskey, C. T., Tompkins, R., Scolnick, E., Caryk, T., & Nirenberg, M. (1968). Sequential translation of trinucleotide codons for the initiation and termination of protein synthesis. *Science, 162*(3849), 135-138.

Chakravorty, S., Helb, D., Burday, M., Connell, N., & Alland, D. (2007). A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *J Microbiol Methods, 69*(2), 330-339. doi:10.1016/j.mimet.2007.02.005

Chang, B., Halgamuge, S., & Tang, S. L. (2006). Analysis of SD sequences in completed microbial genomes: non-SD-led genes are as common as SD-led genes. *Gene, 373*, 90-99. doi:10.1016/j.gene.2006.01.033

Cheadle, C., Fan, J., Cho-Chung, Y. S., Werner, T., Ray, J., Do, L., . . . Becker, K. G. (2005). Stability regulation of mRNA and the control of gene expression. *Ann N Y Acad Sci, 1058*, 196-204. doi:10.1196/annals.1359.026

Chen, H., Bjerknes, M., Kumar, R., & Jay, E. (1994). Determination of the optimal aligned spacing between the Shine-Dalgarno sequence and the translation initiation codon of Escherichia coli mRNAs. *Nucleic Acids Res, 22*(23), 4953-4957.

Clarridge, J. E., 3rd. (2004). Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clin Microbiol Rev, 17*(4), 840-862, table of contents. doi:10.1128/CMR.17.4.840-862.2004

Coghlan, A., & Wolfe, K. H. (2000). Relationship of codon bias to mRNA concentration and protein length in Saccharomyces cerevisiae. *Yeast, 16*(12), 1131-1145. doi:10.1002/1097-0061(20000915)16:12<1131::AID-YEA609>3.0.CO;2-F

Comeron, J. M., & Aguade, M. (1998). An evaluation of measures of synonymous codon usage bias. *J Mol Evol, 47*(3), 268-274.

Cory, S., & Marcker, K. A. (1970). The nucleotide sequence of methionine transfer RNA-M. *Eur J Biochem, 12*(1), 177-194.

de Smit, M. H., & van Duin, J. (1990). Secondary structure of the ribosome binding site determines translational efficiency: a quantitative analysis. *Proc Natl Acad Sci U S A, 87*(19), 7668-7672.
de Smit, M. H., & van Duin, J. (1994). Translational initiation on structured messengers. Another role for the Shine-Dalgarno interaction. *J Mol Biol, 235*(1), 173-184.
Dekel, E., & Alon, U. (2005). Optimality and evolutionary tuning of the expression level of a protein. *Nature, 436*(7050), 588-592. doi:10.1038/nature03842
Dontsova, O., Kopylov, A., & Brimacombe, R. (1991). The location of mRNA in the ribosomal 30S initiation complex; site-directed cross-linking of mRNA analogues carrying several photo-reactive labels simultaneously on either side of the AUG start codon. *EMBO J, 10*(9), 2613-2620.
dos Reis, M., Savva, R., & Wernisch, L. (2004). Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res, 32*(17), 5036-5044. doi:10.1093/nar/gkh834
Dreyfus, M. (1988). What constitutes the signal for the initiation of protein synthesis on Escherichia coli mRNAs? *J Mol Biol, 204*(1), 79-94.
Dunn, J. J., Buzash-Pollert, E., & Studier, F. W. (1978). Mutations of bacteriophage T7 that affect initiation of synthesis of the gene 0.3 protein. *Proc Natl Acad Sci U S A, 75*(6), 2741-2745.
Duret, L., & Mouchiroud, D. (1999). Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. *Proc Natl Acad Sci U S A, 96*(8), 4482-4487.
Duval, M., Korepanov, A., Fuchsbauer, O., Fechter, P., Haller, A., Fabbretti, A., . . . Marzi, S. (2013). Escherichia coli ribosomal protein S1 unfolds structured mRNAs onto the ribosome for active translation initiation. *PLoS Biol, 11*(12), e1001731. doi:10.1371/journal.pbio.1001731
Eckhardt, H., & Luhrmann, R. (1979). Blocking of the initiation of protein biosynthesis by a pentanucleotide complementary to the 3' end of Escherichia coli 16 S rRNA. *J Biol Chem, 254*(22), 11185-11188.
Ehrenberg, M., & Kurland, C. G. (1984). Costs of accuracy determined by a maximal growth rate constraint. *Q Rev Biophys, 17*(1), 45-82.
Emery, L. R., & Sharp, P. M. (2011). Impact of translational selection on codon usage bias in the archaeon Methanococcus maripaludis. *Biol Lett, 7*(1), 131-135. doi:10.1098/rsbl.2010.0620
Espah Borujeni, A., Channarasappa, A. S., & Salis, H. M. (2014). Translation rate is controlled by coupled trade-offs between site accessibility, selective RNA unfolding and sliding at upstream standby sites. *Nucleic Acids Res, 42*(4), 2646-2659. doi:10.1093/nar/gkt1139
Fargo, D. C., Zhang, M., Gillham, N. W., & Boynton, J. E. (1998). Shine-Dalgarno-like sequences are not required for translation of chloroplast mRNAs in Chlamydomonas reinhardtii chloroplasts or in Escherichia coli. *Mol Gen Genet, 257*(3), 271-282.
Farwell, M. A., Roberts, M. W., & Rabinowitz, J. C. (1992). The effect of ribosomal protein S1 from Escherichia coli and Micrococcus luteus on protein synthesis in vitro by E. coli and Bacillus subtilis. *Mol Microbiol, 6*(22), 3375-3383.
Freistroffer, D. V., Pavlov, M. Y., MacDougall, J., Buckingham, R. H., & Ehrenberg, M. (1997). Release factor RF3 in E.coli accelerates the dissociation of release factors RF1 and RF2

from the ribosome in a GTP-dependent manner. *EMBO J, 16*(13), 4126-4133. doi:10.1093/emboj/16.13.4126
Furuichi, Y., & Miura, K. (1975). A blocked structure at the 5' terminus of mRNA from cytoplasmic polyhedrosis virus. *Nature, 253*(5490), 374-375.
Gallie, D. R., & Kado, C. I. (1989). A translational enhancer derived from tobacco mosaic virus is functionally equivalent to a Shine-Dalgarno sequence. *Proc Natl Acad Sci U S A, 86*(1), 129-132.
Giliberti, J., O'Donnell, S., Etten, W. J., & Janssen, G. R. (2012). A 5'-terminal phosphate is required for stable ternary complex formation and translation of leaderless mRNA in Escherichia coli. *RNA, 18*(3), 508-518. doi:10.1261/rna.027698.111
Gold, L. (1988). Posttranscriptional regulatory mechanisms in Escherichia coli. *Annu Rev Biochem, 57*, 199-233. doi:10.1146/annurev.bi.57.070188.001215
Gold, L., Pribnow, D., Schneider, T., Shinedling, S., Singer, B. S., & Stormo, G. (1981). Translational initiation in prokaryotes. *Annu Rev Microbiol, 35*, 365-403. doi:10.1146/annurev.mi.35.100181.002053
Goodman, D. B., Church, G. M., & Kosuri, S. (2013). Causes and effects of N-terminal codon bias in bacterial genes. *Science, 342*(6157), 475-479. doi:10.1126/science.1241934
Gouy, M., & Gautier, C. (1982). Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res, 10*(22), 7055-7074.
Grantham, R., Gautier, C., Gouy, M., Mercier, R., & Pave, A. (1980). Codon catalog usage and the genome hypothesis. *Nucleic Acids Res, 8*(1), r49-r62.
Grill, S., Gualerzi, C. O., Londei, P., & Blasi, U. (2000). Selective stimulation of translation of leaderless mRNA by initiation factor 2: evolutionary implications for translation. *EMBO J, 19*(15), 4101-4110. doi:10.1093/emboj/19.15.4101
Grosjean, H., Marck, C., & de Crecy-Lagard, V. (2007). The various strategies of codon decoding in organisms of the three domains of life: evolutionary implications. *Nucleic Acids Symp Ser (Oxf)*(51), 15-16. doi:10.1093/nass/nrm008
Gu, W., Zhou, T., & Wilke, C. O. (2010). A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. *PLoS Comput Biol, 6*(2), e1000664. doi:10.1371/journal.pcbi.1000664
Gualerzi, C. O., & Pon, C. L. (1990). Initiation of mRNA translation in prokaryotes. *Biochemistry, 29*(25), 5881-5889.
Hansen, P. K., Wikman, F., Clark, B. F., Hershey, J. W., & Uffe Petersen, H. (1986). Interaction between initiator Met-tRNAfMet and elongation factor EF-Tu from E. coli. *Biochimie, 68*(5), 697-703.
Hargrove, J. L., & Schmidt, F. H. (1989). The role of mRNA and protein stability in gene expression. *FASEB J, 3*(12), 2360-2370.
Harold, F. M. (1986). *The vital force : a study of bioenergetics*. New York: W.H. Freeman.
Hartz, D., McPheeters, D. S., & Gold, L. (1991). Influence of mRNA determinants on translation initiation in Escherichia coli. *J Mol Biol, 218*(1), 83-97.
Hershey, J. W., Sonenberg, N., & Mathews, M. B. (2012). Principles of translational control: an overview. *Cold Spring Harb Perspect Biol, 4*(12). doi:10.1101/cshperspect.a011528
Hilterbrand, A., Saelens, J., & Putonti, C. (2012). CBDB: the codon bias database. *BMC Bioinformatics, 13*, 62. doi:10.1186/1471-2105-13-62
Hofacker, I. L. (2003). Vienna RNA secondary structure server. *Nucleic Acids Res, 31*(13), 3429-3431.

Hui, A., & de Boer, H. A. (1987). Specialized ribosome system: preferential translation of a single mRNA species by a subpopulation of mutated ribosomes in Escherichia coli. *Proc Natl Acad Sci U S A, 84*(14), 4762-4766.

Ikemura, T. (1981a). Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes. *J Mol Biol, 146*(1), 1-21.

Ikemura, T. (1981b). Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. *J Mol Biol, 151*(3), 389-409.

Ikemura, T. (1982). Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer RNAs. *J Mol Biol, 158*(4), 573-597.

Isono, K., & Isono, S. (1976). Lack of ribosomal protein S1 in Bacillus stearothermophilus. *Proc Natl Acad Sci U S A, 73*(3), 767-770.

Jacob, W. F., Santer, M., & Dahlberg, A. E. (1987). A single base change in the Shine-Dalgarno region of 16S rRNA of Escherichia coli affects translation of many proteins. *Proc Natl Acad Sci U S A, 84*(14), 4757-4761.

Jacques, N., & Dreyfus, M. (1990). Translation initiation in Escherichia coli: old and new questions. *Mol Microbiol, 4*(7), 1063-1067.

Janda, J. M., & Abbott, S. L. (2007). 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *J Clin Microbiol, 45*(9), 2761-2764. doi:10.1128/JCM.01228-07

Kaminishi, T., Wilson, D. N., Takemoto, C., Harms, J. M., Kawazoe, M., Schluenzen, F., . . . Yokoyama, S. (2007). A snapshot of the 30S ribosomal subunit capturing mRNA via the Shine-Dalgarno interaction. *Structure, 15*(3), 289-297. doi:10.1016/j.str.2006.12.008

Kanaya, S., Yamada, Y., Kudo, Y., & Ikemura, T. (1999). Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene, 238*(1), 143-155.

Karlin, S., Mrazek, J., Ma, J., & Brocchieri, L. (2005). Predicted highly expressed genes in archaeal genomes. *Proc Natl Acad Sci U S A, 102*(20), 7303-7308. doi:10.1073/pnas.0502313102

Kitahara, K., Yasutake, Y., & Miyazaki, K. (2012). Mutational robustness of 16S ribosomal RNA, shown by experimental horizontal gene transfer in Escherichia coli. *Proc Natl Acad Sci U S A, 109*(47), 19220-19225. doi:10.1073/pnas.1213609109

Kolb, A., Hermoso, J. M., Thomas, J. O., & Szer, W. (1977). Nucleic acid helix-unwinding properties of ribosomal protein S1 and the role of S1 in mRNA binding to ribosomes. *Proc Natl Acad Sci U S A, 74*(6), 2379-2383.

Kolbert, C. P., & Persing, D. H. (1999). Ribosomal DNA sequencing as a tool for identification of bacterial pathogens. *Curr Opin Microbiol, 2*(3), 299-305. doi:10.1016/S1369-5274(99)80052-6

Komarova, A. V., Tchufistova, L. S., Dreyfus, M., & Boni, I. V. (2005). AU-rich sequences within 5' untranslated leaders enhance translation and stabilize mRNA in Escherichia coli. *J Bacteriol, 187*(4), 1344-1349. doi:10.1128/JB.187.4.1344-1349.2005

Komarova, A. V., Tchufistova, L. S., Supina, E. V., & Boni, I. V. (2002). Protein S1 counteracts the inhibitory effect of the extended Shine-Dalgarno sequence on translation. *RNA, 8*(9), 1137-1147.

Korostelev, A., Trakhanov, S., Asahara, H., Laurberg, M., Lancaster, L., & Noller, H. F. (2007). Interactions and dynamics of the Shine Dalgarno helix in the 70S ribosome. *Proc Natl Acad Sci U S A, 104*(43), 16840-16843. doi:10.1073/pnas.0707850104

Korostelev, A., Trakhanov, S., Laurberg, M., & Noller, H. F. (2006). Crystal structure of a 70S ribosome-tRNA complex reveals functional interactions and rearrangements. *Cell, 126*(6), 1065-1077. doi:10.1016/j.cell.2006.08.032

Kosuri, S., Goodman, D. B., Cambray, G., Mutalik, V. K., Gao, Y., Arkin, A. P., . . . Church, G. M. (2013). Composability of regulatory sequences controlling transcription and translation in Escherichia coli. *Proc Natl Acad Sci U S A, 110*(34), 14024-14029. doi:10.1073/pnas.1301301110

Kozak, M. (1999). Initiation of translation in prokaryotes and eukaryotes. *Gene, 234*(2), 187-208.

Kramer, P., Gabel, K., Pfeiffer, F., & Soppa, J. (2014). Haloferax volcanii, a prokaryotic species that does not use the Shine Dalgarno mechanism for translation initiation at 5'-UTRs. *PLoS One, 9*(4), e94979. doi:10.1371/journal.pone.0094979

Krishnan, K. M., Van Etten, W. J., 3rd, & Janssen, G. R. (2010). Proximity of the start codon to a leaderless mRNA's 5' terminus is a strong positive determinant of ribosome binding and expression in Escherichia coli. *J Bacteriol, 192*(24), 6482-6485. doi:10.1128/JB.00756-10

Kudla, G., Murray, A. W., Tollervey, D., & Plotkin, J. B. (2009). Coding-sequence determinants of gene expression in Escherichia coli. *Science, 324*(5924), 255-258. doi:10.1126/science.1170160

Kurata, T., Nakanishi, S., Hashimoto, M., Taoka, M., Yamazaki, Y., Isobe, T., & Kato, J. (2015). Novel essential gene Involved in 16S rRNA processing in Escherichia coli. *J Mol Biol, 427*(4), 955-965. doi:10.1016/j.jmb.2014.12.013

Lancaster, L., & Noller, H. F. (2005). Involvement of 16S rRNA nucleotides G1338 and A1339 in discrimination of initiator tRNA. *Mol Cell, 20*(4), 623-632. doi:10.1016/j.molcel.2005.10.006

Laursen, B. S., Sorensen, H. P., Mortensen, K. K., & Sperling-Petersen, H. U. (2005). Initiation of protein synthesis in bacteria. *Microbiol Mol Biol Rev, 69*(1), 101-123. doi:10.1128/MMBR.69.1.101-123.2005

Lecompte, O., Ripp, R., Thierry, J. C., Moras, D., & Poch, O. (2002). Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale. *Nucleic Acids Res, 30*(24), 5382-5390.

Lee, K., Holland-Staley, C. A., & Cunningham, P. R. (1996). Genetic analysis of the Shine-Dalgarno interaction: selection of alternative functional mRNA-rRNA combinations. *RNA, 2*(12), 1270-1285.

Li, G. W., Oh, E., & Weissman, J. S. (2012). The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature, 484*(7395), 538-541. doi:10.1038/nature10965

Li, M., Hu, Q., Xuan, J., Deng, D., & Weng, M. (2003). lambdaN gene expression regulated by translation termination in ribosome L24 mutant. *Sci China C Life Sci, 46*(2), 127-134. doi:10.1360/03yc9014

Liebhaber, S. A. (1997). mRNA stability and the control of gene expression. *Nucleic Acids Symp Ser*(36), 29-32.
Liljenstrom, H., & von Heijne, G. (1987). Translation rate modification by preferential codon usage: intragenic position effects. *J Theor Biol, 124*(1), 43-55.
Lim, K., Furuta, Y., & Kobayashi, I. (2012). Large variations in bacterial ribosomal RNA genes. *Mol Biol Evol, 29*(10), 2937-2948. doi:10.1093/molbev/mss101
Loechel, S., Inamine, J. M., & Hu, P. C. (1991). A novel translation initiation region from Mycoplasma genitalium that functions in Escherichia coli. *Nucleic Acids Res, 19*(24), 6905-6911.
Londei, P. (2005). Evolution of translational initiation: new insights from the archaea. *FEMS Microbiol Rev, 29*(2), 185-200. doi:10.1016/j.femsre.2004.10.002
Londei, P. (2009). Translation Initiation Models in Prokaryotes and Eukaryotes.
Lu, P., Vogel, C., Wang, R., Yao, X., & Marcotte, E. M. (2007). Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol, 25*(1), 117-124. doi:10.1038/nbt1270
Luhrmann, R., Stoffler-Meilicke, M., & Stoffler, G. (1981). Localization of the 3' end of 16S rRNA in Escherichia coli 30S ribosomal subunits by immuno electron microscopy. *Mol Gen Genet, 182*(3), 369-376.
Ma, J., Campbell, A., & Karlin, S. (2002). Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures. *J Bacteriol, 184*(20), 5733-5745.
Malys, N., & McCarthy, J. E. (2011). Translation initiation: variations in the mechanism can be anticipated. *Cell Mol Life Sci, 68*(6), 991-1003. doi:10.1007/s00018-010-0588-z
Marck, C., & Grosjean, H. (2002). tRNomics: analysis of tRNA genes from 50 genomes of Eukarya, Archaea, and Bacteria reveals anticodon-sparing strategies and domain-specific features. *RNA, 8*(10), 1189-1232.
Markham, N. R., & Zuker, M. (2005). DINAMelt web server for nucleic acid melting prediction. *Nucleic Acids Res, 33*(Web Server issue), W577-581. doi:10.1093/nar/gki591
Marquez, V., Frohlich, T., Armache, J. P., Sohmen, D., Donhofer, A., Mikolajka, A., . . . Wilson, D. N. (2011). Proteomic characterization of archaeal ribosomes reveals the presence of novel archaeal-specific ribosomal proteins. *J Mol Biol, 405*(5), 1215-1232. doi:10.1016/j.jmb.2010.11.055
Martin, A. P. (1995). Metabolic rate and directional nucleotide substitution in animal mitochondrial DNA. *Mol Biol Evol, 12*(6), 1124-1131.
McInerney, J. O. (1998). Replicational and transcriptional selection on codon usage in Borrelia burgdorferi. *Proc Natl Acad Sci U S A, 95*(18), 10698-10703.
McLaughlin, J. R., Murray, C. L., & Rabinowitz, J. C. (1981). Unique features in the ribosome binding site sequence of the gram-positive Staphylococcus aureus beta-lactamase gene. *J Biol Chem, 256*(21), 11283-11291.
Melancon, P., Leclerc, D., Destroismaisons, N., & Brakier-Gingras, L. (1990). The anti-Shine-Dalgarno region in Escherichia coli 16S ribosomal RNA is not essential for the correct selection of translational starts. *Biochemistry, 29*(13), 3402-3407.
Milo, R., Jorgensen, P., Moran, U., Weber, G., & Springer, M. (2010). BioNumbers--the database of key numbers in molecular and cell biology. *Nucleic Acids Res, 38*(Database issue), D750-753. doi:10.1093/nar/gkp889

Milon, P., Maracci, C., Filonava, L., Gualerzi, C. O., & Rodnina, M. V. (2012). Real-time assembly landscape of bacterial 30S translation initiation complex. *Nat Struct Mol Biol, 19*(6), 609-615. doi:10.1038/nsmb.2285

Milon, P., & Rodnina, M. V. (2012). Kinetic control of translation initiation in bacteria. *Crit Rev Biochem Mol Biol, 47*(4), 334-348. doi:10.3109/10409238.2012.678284

Moriyama, E. N., & Powell, J. R. (1997). Codon usage bias and tRNA abundance in Drosophila. *J Mol Evol, 45*(5), 514-523.

Murray, C. L., & Rabinowitz, J. C. (1982). Nucleotide sequences of transcription and translation initiation regions in Bacillus phage phi 29 early genes. *J Biol Chem, 257*(2), 1053-1062.

Mutalik, V. K., Guimaraes, J. C., Cambray, G., Mai, Q. A., Christoffersen, M. J., Martin, L., . . . Arkin, A. P. (2013). Quantitative estimation of activity and quality for collections of functional genetic elements. *Nat Methods, 10*(4), 347-353. doi:10.1038/nmeth.2403

Muto, A., & Osawa, S. (1987). The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc Natl Acad Sci U S A, 84*(1), 166-169.

Myasnikov, A. G., Simonetti, A., Marzi, S., & Klaholz, B. P. (2009). Structure-function insights into prokaryotic and eukaryotic translation initiation. *Curr Opin Struct Biol, 19*(3), 300-309. doi:10.1016/j.sbi.2009.04.010

Na, D., & Lee, D. (2010). RBSDesigner: software for designing synthetic ribosome binding sites that yields a desired level of protein expression. *Bioinformatics, 26*(20), 2633-2634. doi:10.1093/bioinformatics/btq458

Nakagawa, S., Niimura, Y., Miura, K., & Gojobori, T. (2010). Dynamic evolution of translation initiation mechanisms in prokaryotes. *Proc Natl Acad Sci U S A, 107*(14), 6382-6387. doi:10.1073/pnas.1002036107

Nakamoto, T. (2006). A unified view of the initiation of protein synthesis. *Biochem Biophys Res Commun, 341*(3), 675-678. doi:10.1016/j.bbrc.2006.01.019

Nivinskas, R., Malys, N., Klausa, V., Vaiskunaite, R., & Gineikiene, E. (1999). Post-transcriptional control of bacteriophage T4 gene 25 expression: mRNA secondary structure that enhances translational initiation. *J Mol Biol, 288*(3), 291-304. doi:10.1006/jmbi.1999.2695

Noah, J. W., Shapkina, T., & Wollenzien, P. (2000). UV-induced crosslinks in the 16S rRNAs of Escherichia coli, Bacillus subtilis and Thermus aquaticus and their implications for ribosome structure and photochemistry. *Nucleic Acids Res, 28*(19), 3785-3792.

O'Donnell, S. M., & Janssen, G. R. (2001). The initiation codon affects ribosome binding and translational efficiency in Escherichia coli of cI mRNA with or without the 5' untranslated leader. *J Bacteriol, 183*(4), 1277-1283. doi:10.1128/JB.183.4.1277-1283.2001

O'Donnell, S. M., & Janssen, G. R. (2002). Leaderless mRNAs bind 70S ribosomes more strongly than 30S ribosomal subunits in Escherichia coli. *J Bacteriol, 184*(23), 6730-6733.

Ogle, J. M., Brodersen, D. E., Clemons, W. M., Jr., Tarry, M. J., Carter, A. P., & Ramakrishnan, V. (2001). Recognition of cognate transfer RNA by the 30S ribosomal subunit. *Science, 292*(5518), 897-902. doi:10.1126/science.1060612

Olsthoorn, R. C., Zoog, S., & van Duin, J. (1995). Coevolution of RNA helix stability and Shine-Dalgarno complementarity in a translational start region. *Mol Microbiol, 15*(2), 333-339.

Orso, S., Gouy, M., Navarro, E., & Normand, P. (1994). Molecular phylogenetic analysis of Nitrobacter spp. *Int J Syst Bacteriol, 44*(1), 83-86. doi:10.1099/00207713-44-1-83

Osada, Y., Saito, R., & Tomita, M. (1999). Analysis of base-pairing potentials between 16S rRNA and 5' UTR for translation initiation in various prokaryotes. *Bioinformatics, 15*(7-8), 578-581.

Osterman, I. A., Evfratov, S. A., Sergiev, P. V., & Dontsova, O. A. (2013). Comparison of mRNA features affecting translation initiation and reinitiation. *Nucleic Acids Res, 41*(1), 474-486. doi:10.1093/nar/gks989

Pace, N. R. (1997). A molecular view of microbial diversity and the biosphere. *Science, 276*(5313), 734-740.

Palys, T., Nakamura, L. K., & Cohan, F. M. (1997). Discovery and classification of ecological diversity in the bacterial world: the role of DNA sequence data. *Int J Syst Bacteriol, 47*(4), 1145-1156. doi:10.1099/00207713-47-4-1145

Patel, J. B. (2001). 16S rRNA gene sequencing for bacterial pathogen identification in the clinical laboratory. *Mol Diagn, 6*(4), 313-321. doi:10.1054/modi.2001.29158

Poole, E. S., Brown, C. M., & Tate, W. P. (1995). The identity of the base following the stop codon determines the efficiency of in vivo translational termination in Escherichia coli. *EMBO J, 14*(1), 151-158.

Prabhakaran, R., Chithambaram, S., & Xia, X. (2015). Escherichia coli and Staphylococcus phages: effect of translation initiation efficiency on differential codon adaptation mediated by virulent and temperate lifestyles. *J Gen Virol, 96*(Pt 5), 1169-1179. doi:10.1099/vir.0.000050

Qin, D., Abdi, N. M., & Fredrick, K. (2007). Characterization of 16S rRNA mutations that decrease the fidelity of translation initiation. *RNA, 13*(12), 2348-2355. doi:10.1261/rna.715307

Qin, D., & Fredrick, K. (2009). Control of translation initiation involves a factor-induced rearrangement of helix 44 of 16S ribosomal RNA. *Mol Microbiol, 71*(5), 1239-1249. doi:10.1111/j.1365-2958.2009.06598.x

Qing, G., Xia, B., & Inouye, M. (2003). Enhancement of translation initiation by A/T-rich sequences downstream of the initiation codon in Escherichia coli. *J Mol Microbiol Biotechnol, 6*(3-4), 133-144. doi:77244

Qu, X., Lancaster, L., Noller, H. F., Bustamante, C., & Tinoco, I., Jr. (2012). Ribosomal protein S1 unwinds double-stranded RNA in multiple steps. *Proc Natl Acad Sci U S A, 109*(36), 14458-14463. doi:10.1073/pnas.1208950109

Reisbig, M. D., & Hanson, N. D. (2004). Promoter sequences necessary for high-level expression of the plasmid-associated ampC beta-lactamase gene blaMIR-1. *Antimicrob Agents Chemother, 48*(11), 4177-4182. doi:10.1128/AAC.48.11.4177-4182.2004

Ringquist, S., Shinedling, S., Barrick, D., Green, L., Binkley, J., Stormo, G. D., & Gold, L. (1992). Translation initiation in Escherichia coli: sequences within the ribosome-binding site. *Mol Microbiol, 6*(9), 1219-1229.

Rinke-Appel, J., Junke, N., Brimacombe, R., Lavrik, I., Dokudovskaya, S., Dontsova, O., & Bogdanov, A. (1994). Contacts between 16S ribosomal RNA and mRNA, within the spacer region separating the AUG initiator codon and the Shine-Dalgarno sequence; a site-directed cross-linking study. *Nucleic Acids Res, 22*(15), 3018-3025.

Roberts, M. W., & Rabinowitz, J. C. (1989). The effect of Escherichia coli ribosomal protein S1 on the translational specificity of bacterial ribosomes. *J Biol Chem, 264*(4), 2228-2235.

Robinson, M., Lilley, R., Little, S., Emtage, J. S., Yarranton, G., Stephens, P., . . . Humphreys, G. (1984). Codon usage can affect efficiency of translation of genes in Escherichia coli. *Nucleic Acids Res, 12*(17), 6663-6671.

Rocha, E. P., & Danchin, A. (2002). Base composition bias might result from competition for metabolic resources. *Trends Genet, 18*(6), 291-294. doi:10.1016/S0168-9525(02)02690-2

Sacerdot, C., Chiaruttini, C., Engst, K., Graffe, M., Milet, M., Mathy, N., . . . Springer, M. (1996). The role of the AUU initiation codon in the negative feedback regulation of the gene for translation initiation factor IF3 in Escherichia coli. *Mol Microbiol, 21*(2), 331-346.

Salah, P., Bisaglia, M., Aliprandi, P., Uzan, M., Sizun, C., & Bontems, F. (2009). Probing the relationship between Gram-negative and Gram-positive S1 proteins by sequence analysis. *Nucleic Acids Res, 37*(16), 5578-5588. doi:10.1093/nar/gkp547

Salis, H. M., Mirsky, E. A., & Voigt, C. A. (2009). Automated design of synthetic ribosome binding sites to control protein expression. *Nat Biotechnol, 27*(10), 946-950. doi:10.1038/nbt.1568

Sartorius-Neef, S., & Pfeifer, F. (2004). In vivo studies on putative Shine-Dalgarno sequences of the halophilic archaeon Halobacterium salinarum. *Mol Microbiol, 51*(2), 579-588. doi:10.1046/j.1365-2958.2003.03858.x

Scharff, L. B., Childs, L., Walther, D., & Bock, R. (2011). Local absence of secondary structure permits translation of mRNAs that lack ribosome-binding sites. *PLoS Genet, 7*(6), e1002155. doi:10.1371/journal.pgen.1002155

Schneider, T. D., Stormo, G. D., Gold, L., & Ehrenfeucht, A. (1986). Information content of binding sites on nucleotide sequences. *J Mol Biol, 188*(3), 415-431.

Schurr, T., Nadir, E., & Margalit, H. (1993). Identification and characterization of E.coli ribosomal binding sites by free energy computation. *Nucleic Acids Res, 21*(17), 4019-4023.

Schwanhausser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., . . . Selbach, M. (2011). Global quantification of mammalian gene expression control. *Nature, 473*(7347), 337-342. doi:10.1038/nature10098

Scolnick, E., Tompkins, R., Caskey, T., & Nirenberg, M. (1968). Release factors differing in specificity for terminator codons. *Proc Natl Acad Sci U S A, 61*(2), 768-774.

Scolnick, E. M., & Caskey, C. T. (1969). Peptide chain termination. V. The role of release factors in mRNA terminator codon recognition. *Proc Natl Acad Sci U S A, 64*(4), 1235-1241.

Selmer, M., Dunham, C. M., Murphy, F. V. t., Weixlbaumer, A., Petry, S., Kelley, A. C., . . . Ramakrishnan, V. (2006). Structure of the 70S ribosome complexed with mRNA and tRNA. *Science, 313*(5795), 1935-1942. doi:10.1126/science.1131127

Sengupta, J., Agrawal, R. K., & Frank, J. (2001). Visualization of protein S1 within the 30S ribosomal subunit and its interaction with messenger RNA. *Proc Natl Acad Sci U S A, 98*(21), 11991-11996. doi:10.1073/pnas.211266898

Sharp, P. M., Bailes, E., Grocock, R. J., Peden, J. F., & Sockett, R. E. (2005). Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Res, 33*(4), 1141-1153. doi:10.1093/nar/gki242

Sharp, P. M., & Devine, K. M. (1989). Codon usage and gene expression level in Dictyostelium discoideum: highly expressed genes do 'prefer' optimal codons. *Nucleic Acids Res, 17*(13), 5029-5039.

Sharp, P. M., & Li, W. H. (1987). The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res, 15*(3), 1281-1295.
Shean, C. S., & Gottesman, M. E. (1992). Translation of the prophage lambda cl transcript. *Cell, 70*(3), 513-522.
Shell, S. S., Wang, J., Lapierre, P., Mir, M., Chase, M. R., Pyle, M. M., . . . Gray, T. A. (2015). Leaderless Transcripts and Small Proteins Are Common Features of the Mycobacterial Translational Landscape. *PLoS Genet, 11*(11), e1005641. doi:10.1371/journal.pgen.1005641
Shine, J., & Dalgarno, L. (1974). The 3'-terminal sequence of Escherichia coli 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc Natl Acad Sci U S A, 71*(4), 1342-1346.
Simonetti, A., Marzi, S., Jenner, L., Myasnikov, A., Romby, P., Yusupova, G., . . . Yusupov, M. (2009). A structural view of translation initiation in bacteria. *Cell Mol Life Sci, 66*(3), 423-436. doi:10.1007/s00018-008-8416-4
Slupska, M. M., King, A. G., Fitz-Gibbon, S., Besemer, J., Borodovsky, M., & Miller, J. H. (2001). Leaderless transcripts of the crenarchaeal hyperthermophile Pyrobaculum aerophilum. *J Mol Biol, 309*(2), 347-360. doi:10.1006/jmbi.2001.4669
Sorensen, M. A., Fricke, J., & Pedersen, S. (1998). Ribosomal protein S1 is required for translation of most, if not all, natural mRNAs in Escherichia coli in vivo. *J Mol Biol, 280*(4), 561-569. doi:10.1006/jmbi.1998.1909
Sorensen, M. A., Kurland, C. G., & Pedersen, S. (1989). Codon usage determines translation rate in Escherichia coli. *J Mol Biol, 207*(2), 365-377.
Sorokin, A., Serror, P., Pujic, P., Azevedo, V., & Ehrlich, S. D. (1995). The Bacillus subtilis chromosome region encoding homologues of the Escherichia coli mssA and rpsA gene products. *Microbiology, 141 (Pt 2)*, 311-319. doi:10.1099/13500872-141-2-311
Starmer, J., Stomp, A., Vouk, M., & Bitzer, D. (2006). Predicting Shine-Dalgarno sequence locations exposes genome annotation errors. *PLoS Comput Biol, 2*(5), e57. doi:10.1371/journal.pcbi.0020057
Steitz, J. A., & Jakes, K. (1975). How ribosomes select initiator regions in mRNA: base pair formation between the 3' terminus of 16S rRNA and the mRNA during initiation of protein synthesis in Escherichia coli. *Proc Natl Acad Sci U S A, 72*(12), 4734-4738.
Stothard, P., Van Domselaar, G., Shrivastava, S., Guo, A., O'Neill, B., Cruz, J., . . . Wishart, D. S. (2005). BacMap: an interactive picture atlas of annotated bacterial genomes. *Nucleic Acids Res, 33*(Database issue), D317-320. doi:10.1093/nar/gki075
Supek, F., & Smuc, T. (2010). On relevance of codon usage to expression of synthetic and natural genes in Escherichia coli. *Genetics, 185*(3), 1129-1134. doi:10.1534/genetics.110.115477
Sussman, J. K., Simons, E. L., & Simons, R. W. (1996). Escherichia coli translation initiation factor 3 discriminates the initiation codon in vivo. *Mol Microbiol, 21*(2), 347-360.
Taniguchi, T., & Weissmann, C. (1978). Inhibition of Qbeta RNA 70S ribosome initiation complex formation by an oligonucleotide complementary to the 3' terminal region of E. coli 16S ribosomal RNA. *Nature, 275*(5682), 770-772.
Taniguchi, Y., Choi, P. J., Li, G. W., Chen, H., Babu, M., Hearn, J., . . . Xie, X. S. (2010). Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science, 329*(5991), 533-538. doi:10.1126/science.1188308

Tedin, K., Resch, A., & Blasi, U. (1997). Requirements for ribosomal protein S1 for translation initiation of mRNAs with and without a 5' leader sequence. *Mol Microbiol, 25*(1), 189-199.

Tegel, H., Ottosson, J., & Hober, S. (2011). Enhancing the protein production levels in Escherichia coli with a strong promoter. *FEBS J, 278*(5), 729-739. doi:10.1111/j.1742-4658.2010.07991.x

Tolstrup, N., Sensen, C. W., Garrett, R. A., & Clausen, I. G. (2000). Two different and highly organized mechanisms of translation initiation in the archaeon Sulfolobus solfataricus. *Extremophiles, 4*(3), 175-179.

Tortoli, E. (2003). Impact of genotypic studies on mycobacterial taxonomy: the new mycobacteria of the 1990s. *Clin Microbiol Rev, 16*(2), 319-354.

Tu, C., Zhou, X., Tropea, J. E., Austin, B. P., Waugh, D. S., Court, D. L., & Ji, X. (2009). Structure of ERA in complex with the 3' end of 16S rRNA: implications for ribosome biogenesis. *Proc Natl Acad Sci U S A, 106*(35), 14843-14848. doi:10.1073/pnas.0904032106

Tuller, T., Waldman, Y. Y., Kupiec, M., & Ruppin, E. (2010). Translation efficiency is determined by both codon bias and folding energy. *Proc Natl Acad Sci U S A, 107*(8), 3645-3650. doi:10.1073/pnas.0909910107

Tzareva, N. V., Makhno, V. I., & Boni, I. V. (1994). Ribosome-messenger recognition in the absence of the Shine-Dalgarno interactions. *FEBS Lett, 337*(2), 189-194.

Van Etten, W. J., & Janssen, G. R. (1998). An AUG initiation codon, not codon-anticodon complementarity, is required for the translation of unleadered mRNA in Escherichia coli. *Mol Microbiol, 27*(5), 987-1001.

Vanin, E. F. (1985). Processed pseudogenes: characteristics and evolution. *Annu Rev Genet, 19*, 253-272. doi:10.1146/annurev.ge.19.120185.001345

Vellanoweth, R. L., & Rabinowitz, J. C. (1992). The influence of ribosome-binding-site elements on translational efficiency in Bacillus subtilis and Escherichia coli in vivo. *Mol Microbiol, 6*(9), 1105-1114.

Vesper, O., Amitai, S., Belitsky, M., Byrgazov, K., Kaberdina, A. C., Engelberg-Kulka, H., & Moll, I. (2011). Selective translation of leaderless mRNAs by specialized ribosomes generated by MazF in Escherichia coli. *Cell, 147*(1), 147-157. doi:10.1016/j.cell.2011.07.047

Vimberg, V., Tats, A., Remm, M., & Tenson, T. (2007). Translation initiation region sequence preferences in Escherichia coli. *BMC Mol Biol, 8*, 100. doi:10.1186/1471-2199-8-100

Vishwanath, P., Favaretto, P., Hartman, H., Mohr, S. C., & Smith, T. F. (2004). Ribosomal protein-sequence block structure suggests complex prokaryotic evolution with implications for the origin of eukaryotes. *Mol Phylogenet Evol, 33*(3), 615-625. doi:10.1016/j.ympev.2004.07.003

Vogel, C., Abreu Rde, S., Ko, D., Le, S. Y., Shapiro, B. A., Burns, S. C., . . . Penalva, L. O. (2010). Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Mol Syst Biol, 6*, 400. doi:10.1038/msb.2010.59

Vogel, C., & Marcotte, E. M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat Rev Genet, 13*(4), 227-232. doi:10.1038/nrg3185

Wang, M., Weiss, M., Simonovic, M., Haertinger, G., Schrimpf, S. P., Hengartner, M. O., & von Mering, C. (2012). PaxDb, a database of protein abundance averages across all three domains of life. *Mol Cell Proteomics, 11*(8), 492-500. doi:10.1074/mcp.O111.014704

Wilson, T. M. (1986). Expression of the large 5'-proximal cistron of tobacco mosaic virus by 70 S ribosomes during cotranslational disassembly in a prokaryotic cell-free system. *Virology, 152*(1), 277-279.

Woese, C. R. (1987). Bacterial evolution. *Microbiol Rev, 51*(2), 221-271.

Woese, C. R., Magrum, L. J., Gupta, R., Siegel, R. B., Stahl, D. A., Kop, J., . . . Noller, H. F. (1980). Secondary structure model for bacterial 16S ribosomal RNA: phylogenetic, enzymatic and chemical evidence. *Nucleic Acids Res, 8*(10), 2275-2293.

Woese, C. R., Stackebrandt, E., Macke, T. J., & Fox, G. E. (1985). A phylogenetic definition of the major eubacterial taxa. *Syst Appl Microbiol, 6*, 143-151.

Wright, F. (1990). The 'effective number of codons' used in a gene. *Gene, 87*(1), 23-29.

Wu, C. J., & Janssen, G. R. (1997). Expression of a streptomycete leaderless mRNA encoding chloramphenicol acetyltransferase in Escherichia coli. *J Bacteriol, 179*(21), 6824-6830.

Wu, X. Q., Iyengar, P., & RajBhandary, U. L. (1996). Ribosome-initiator tRNA complex as an intermediate in translation initiation in Escherichia coli revealed by use of mutant initiator tRNAs and specialized ribosomes. *EMBO J, 15*(17), 4734-4739.

Wu, X. Q., & RajBhandary, U. L. (1997). Effect of the amino acid attached to Escherichia coli initiator tRNA on its affinity for the initiation factor IF2 and on the IF2 dependence of its binding to the ribosome. *J Biol Chem, 272*(3), 1891-1895.

Xia, X. (1995). Body temperature, rate of biosynthesis, and evolution of genome size. *Mol Biol Evol, 12*(5), 834-842.

Xia, X. (1996). Maximizing transcription efficiency causes codon usage bias. *Genetics, 144*(3), 1309-1320.

Xia, X. (1998). How optimized is the translational machinery in Escherichia coli, Salmonella typhimurium and Saccharomyces cerevisiae? *Genetics, 149*(1), 37-44.

Xia, X. (2007). An improved implementation of codon adaptation index. *Evol Bioinform Online, 3*, 53-58.

Xia, X. (2008). The cost of wobble translation in fungal mitochondrial genomes: integration of two traditional hypotheses. *BMC Evol Biol, 8*, 211. doi:10.1186/1471-2148-8-211

Xia, X. (2012). DNA replication and strand asymmetry in prokaryotic and mitochondrial genomes. *Curr Genomics, 13*(1), 16-27. doi:10.2174/138920212799034776

Xia, X. (2013). DAMBE5: a comprehensive software package for data analysis in molecular biology and evolution. *Mol Biol Evol, 30*(7), 1720-1728. doi:10.1093/molbev/mst064

Xia, X. (2015). A major controversy in codon-anticodon adaptation resolved by a new codon usage index. *Genetics, 199*(2), 573-579. doi:10.1534/genetics.114.172106

Xia, X., & Holcik, M. (2009). Strong eukaryotic IRESs have weak secondary structure. *PLoS One, 4*(1), e4136. doi:10.1371/journal.pone.0004136

Xia, X., Huang, H., Carullo, M., Betran, E., & Moriyama, E. N. (2007). Conflict between translation initiation and elongation in vertebrate mitochondrial genomes. *PLoS One, 2*(2), e227. doi:10.1371/journal.pone.0000227

Xia, X., MacKay, V., Yao, X., Wu, J., Miura, F., Ito, T., & Morris, D. R. (2011). Translation initiation: a regulatory role for poly(A) tracts in front of the AUG codon in Saccharomyces cerevisiae. *Genetics, 189*(2), 469-478. doi:10.1534/genetics.111.132068

Xia, X., Wang, H., Xie, Z., Carullo, M., Huang, H., & Hickey, D. (2006). Cytosine usage modulates the correlation between CDS length and CG content in prokaryotic genomes. *Mol Biol Evol, 23*(7), 1450-1454. doi:10.1093/molbev/msl012

Yao, S., Blaustein, J. B., & Bechhofer, D. H. (2007). Processing of Bacillus subtilis small cytoplasmic RNA: evidence for an additional endonuclease cleavage site. *Nucleic Acids Res, 35*(13), 4464-4473. doi:10.1093/nar/gkm460

Yassin, A., Fredrick, K., & Mankin, A. S. (2005). Deleterious mutations in small subunit ribosomal RNA identify functional sites and potential targets for antibiotics. *Proc Natl Acad Sci U S A, 102*(46), 16620-16625. doi:10.1073/pnas.0508444102

Yusupov, M. M., Yusupova, G. Z., Baucom, A., Lieberman, K., Earnest, T. N., Cate, J. H., & Noller, H. F. (2001). Crystal structure of the ribosome at 5.5 A resolution. *Science, 292*(5518), 883-896. doi:10.1126/science.1060089

Yusupova, G., Jenner, L., Rees, B., Moras, D., & Yusupov, M. (2006). Structural basis for messenger RNA movement on the ribosome. *Nature, 444*(7117), 391-394. doi:10.1038/nature05281

Zavialov, A. V., Mora, L., Buckingham, R. H., & Ehrenberg, M. (2002). Release of peptide promoted by the GGQ motif of class 1 release factors regulates the GTPase activity of RF3. *Mol Cell, 10*(4), 789-798.

Zhang, J. R., & Deutscher, M. P. (1989). Analysis of the upstream region of the Escherichia coli rnd gene encoding RNase D. Evidence for translational regulation of a putative tRNA processing enzyme. *J Biol Chem, 264*(30), 18228-18233.

Zhang, Z., Li, J., Cui, P., Ding, F., Li, A., Townsend, J. P., & Yu, J. (2012). Codon Deviation Coefficient: a novel measure for estimating codon usage bias and its statistical significance. *BMC Bioinformatics, 13*, 43. doi:10.1186/1471-2105-13-43

Zheng, X., Hu, G. Q., She, Z. S., & Zhu, H. (2011). Leaderless genes in bacteria: clue to the evolution of translation initiation mechanisms in prokaryotes. *BMC Genomics, 12*, 361. doi:10.1186/1471-2164-12-361

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