Factors Involved in the Codon Usage Bias Among Different Genes in a Genome, And Among Different Sites Within a Gene

By
Arash Ahmadi

A Thesis
in Partial Fulfilment of the Requirements
for the Degree
Master of Science

McMaster University

Abstract

In this study we have focused on the codon usage bias in *E. coli*, at two different levels, the codon usage bias among the genes in the genome and the codon usage bias among different sites within one gene.

In chapter 3, we use the population genetics model and the data available on the protein and mRNA levels of the *E. coli* genes to understand the pattern of codon usage in different genes with different expression levels. Here, by using likelihood-based statistical tests, we can compare the models with different measures of expression (i.e. total number of proteins produced per cell cycle for each gene, number of mRNA molecules transcribed per cell cycle for each gene, number of proteins produced per mRNA & protein production rate over each mRNA) and see which one best explains the pattern we observe. We also provide an analytic model of protein production in order to further clarify the existence of codon bias in spite of translation being initiation limited and also why the codon bias is observed to be more correlated with total protein level of a gene compared to other measures of expression. Besides codon bias, we are able to test for the existence of context dependent mutation. Our model uses two parameter, a frequency in absence of selection and a selection coefficient, for each codon and by testing the over-parametrization of the model we can see whether only considering the third nucleotide position of the codons, or considering the first two positions, would be sufficient to fit the real data with the model or we have to consider all three nucleotide

positions in codons for finding the most suited frequencies. We have also fitted the model for the codon usage patter in the Yeast and also tested for the context dependent mutation in this organism.

In chapter 4, we focus on the first 10-15 codons in the genes of *E. coli*. Motivated by the fact that in this region we observe two phenomena, reduction in translation efficiency and suppression of mRNA secondary structures, we investigate whether the former is a side effect of selection for the latter. For this matter we have generated a set of synonymous randomized sequences, and then by selecting the ones which show weak secondary structures in the mentioned region, we would be able to test the theory. We will also look at the frequencies of the amino acids in *E. coli* genes and see whether the selection for weak secondary structures in the translation initiation region could be strong enough to not only affect the codon usage, but also the choice of amino acids. We would also provide information on the correlation between the strength of the mRNA secondary structure in the first 13 codons and the overall translation efficiency of the genes.

Acknowledgments

I would like to thank Dr. Paul Higgs for all of his guidance and help without which this work could not be done.

I would also like to thank all my family and friends for their emotional support.

Table of Contents

List of Figures

List of Tables

Chapter 1
Introduction

1.1- The Genetic Code

Soon after the structure of the DNA was discovered in 1953 by Watson & Crick, several attempts started in order to understand how the proteins are translated from the DNA sequence with the four nucleotides (adenine, A, cytosine, C, thymine, T, and guanine, G). George Gamow's suggestion (Crick, 1988) that dividing the DNA sequence into units of three nucleotides would result in the minimum number of translation units, $4^3 = 64$, in order for the cell to translate the 20 amino acids, helped the scientist to encrypt the genetic code and discover what amino acid each codon, the triplets of nucleotides in the DNA sequence, codes for.

Nirenberg and Matthaei were the first to elucidate the nature of a codon in 1961, when they synthesized an mRNA sequence of only including uracil nucleotides (i.e., UUUUUU…) *in vitro* and realized the translated polypeptide contains only phenylalanine (Nirenberg et al., 1961). Successive works done by Severo Ochoa's research group (Lengyel et al., 1961; Speyer et al., 1962; Lengyel et al., 1962), Har Gobind Khorana (1966) and Robert W. Holley (1965) shed more light on our understanding of the genetic code and the protein translation process in the cells.

Not long after *E. coli*'s genetic code was decrypted (Nirenberg et al., 1963), it was suggested that the genetic code, with minor modifications, is universal (Hinegardner & Engelberg, 1963; Woese et al., 1964), which gave it the name "standard code", and also that the assignment of codons to amino acids is not random (Woese, 1965; Crick, 1968). However now with the capacity of sequencing of complete genomes of various species, clear evidence has been provided that there are deviations from the standard code (Knigh et al., 2001 a; Yokobori et al., 2001), and the standard genetic code is not as universal as initially thought (Sengupta et al., 2007).

Several studies show that the position of amino acids in the genetic code is affected by biosynthethic parameters, and the amino acids which have similar biochemical and physicochemical properties tend to have similar codons (Wong, 1975; Amirnovin, 1997; Taylor & Coates, 1989; Giulio, 1997). Such patterns in the arrangement of the amino acids in the genetic code might be due to selection for the codes, in the competition between organisms which showed much different genetic codes in early stages of life on Earth, which would be more robust against potential errors in the translation of the DNA sequence or the single-point mutations in DNA replication (Alff-Steinberger, 1969; Woese, 1973; Haig & Hurst, 1991; Higgs, 2009). By considering these facts and the bias in the mutation rates between the four nucleotide bases, A, T, G & C, Freeland and Hurst (1998) have compared the natural genetic code with a sample of 1 million random genetic codes, by randomly assigning the amino acids to the 64 codons, and have concluded that in terms of robustness against mistranslation and point mutations, only 1 code in that sample shows higher efficiency. In the same paper Freeland and Hurst have argued that selection

for reducing the effects of mistranslation, rather than single-point mutations, might have played a more important role in shaping the current pattern in assigning the amino acids to codons.

But another feature that can be easily noticed by looking at the genetic code is its redundancy. There are 64 triplets, codons, and only 20 amino acids to be coded for. There are also three codons that are coded as translation termination, UAA, UAG & UGA. This gives 41 codons to be distributed between the amino acids. This distribution also is not random and amino acids are coded by 1-6 codons. The codons which code for the same amino acid, and thus do not affect the sequence or function of the translated protein, are called synonymous. This phenomenon has puzzled scientists for a long time to understand the effect of synonymous mutations, the mutation which changes a codon into another one which is synonymous to it. This matter becomes more complicated when we observe that the frequencies of synonymous codons in different genomes or different genes within each genome, are far from being random. Which gives rise to the term "codon-usage bias", or "codon bias" for short.

1.2- Observation of the codon bias:

Codon bias among different species:

Since early 80's, it was observed that despite the fact that different organisms generally share the same genetic code, the direction in the bias between synonymous codons varies between species. These observations, added to the fact that the bias in appearance of the synonymous codons is more or less consistent across most the genes in a genome (Grantham, 1980; Grantham et al., 1980; Ikemura, 1985; Chen et al., 2004), have

led to the “Genome Hypothesis”. Acccording to this hypothesis different organisms have specific codon biases distinguishable from other organisms (Grantham et al., 1980). Besides, by comparing the codon frequencies observed in different organisms (Andersson & Sharp 1996 a, Andersson & Sharp 1996 b), it can be noticed that the strength of this bias also varies between different organisms. One strong factor which can be used to predict the codon bias between different species is the genomic GC content, the fraction of the two nucleotides guanine and cytosine in the genome (Plotkin & Kudla, 2010). In fact, the codon bias variations among different bacterial genomes can be accurately predicted by measuring the nucleotide content of the regions outside the open reading frame (ORF) (Hershberg & Petrov, 2008; Chen et al., 2004; Knight et al., 2001 b).

Codon bias among the genes within a genome:

At the same time it was also observed that in *E. coli*, *Salmonella typhimurium* and *Saccharomyces cerevisiae*, in a subset of the genes within each genome which are highly expressed, the strength and, for some amino acids, the choice of the abundant codon differs significantly from the rest of the genes (Grantham et al., 1981; Ikemura, 1981; Bennetzen & Hall, 1982; Gouy & Gautier, 1982). Comparison of the results from more recent experiments in broader groups of species (Duret, 2002; Duret & Mouchiroud, 1999; Sharp & Li, 1987; Bulmer, 1991; Ran & Higgs, 2010; Eyre-Walker & Bulmer, 1995), reveals the same phenomenon within the genomes (Plotkin & Kudla, 2010).

Codon bias in different positions within a gene:

Even in choosing synonymous codons for different positions in one gene some deviations from randomness can be observed. Studies show a strong deviation from null hypothesis in synonymous codon substitutions in the beginning region of the genes in diverse organisms such as bacteria, yeast and fruit flies (Bulmer, 1988; Qin et al., 2004; Bentele et al., 2013; Tuller et al., 2010). When looking at this region we observe that there's a tendency for choosing the so called inefficient codons, the codons which are thought to be not recognized and translated at high speed.

At the same time it has been shown that a trend of reduction in the strength of the mRNA secondary structure and also in the GC content of the codons in the translation initiation region of the genes in diverse organisms in prokaryotes and eukaryotes exists (Bettany et al., 1989; de Smit & Van Duin, 1990; Gu et al., 2010; Kudla et al., 2009).

1.3- Measures of Codon Bias:

As soon as the codon usage bias was discovered, measures for comparing the strength of codon bias among the species and the genes began to be proposed. Different approaches have been proposed using different statistical methods and different features associated with the patterns observed in the frequency of synonymous codons.

One way of approaching this issue is to work out a measure to see how much the codon frequencies deviate from a postulated unbiased pattern of usage. The method proposed by McLachlan et al. (McLachlan et al., 1984), follows such procedure. Calculating the chi squared value for the deviation from random codon usage has also been

used for measuring the strength of codon bias (Sharp et al., 1986). Ikemura has focused on the relation between translation efficiency and codon bias, and has tried to identify the "optimal" codon among the codons coding for one specific amino acid. Then by calculating the frequency of this optimal codon in the genes, the strength of codon bias can be compared among the genes (Ikemura, 1985). This method would divide the synonymous codons into two groups of "optimal" and "non-optimal".

Gribskov et al. (1984), suggested an index which is based on the ratio of the likelihood of observing a particular codon in a highly expressed gene to the likelihood of finding that codon in a random sequence with the same base composition as that in the sequence under study.

The famous measure of "Codon Adaptation Index", or CAI for short, was introduced in 1987 (Sharp & Li, 1987), which has been referred to by different authors for comparing the extent of codon bias among species and genes. They also focus on the relation between synonymous substitution of the codons and translation efficiency. They introduce a method so that the codons are not just considered as only optimal or non-optimal, but there would be a way of ranking the codons in terms of translation efficiency. Considering the fact that the strength of codon bias is fairly high in some genes, and the correlation between this strength and the expression level of these genes, they introduce a "reference set" of genes which are highly expressed and thought to be under selection to show a strong bias in codon usage, and the codon adaptation index of each codon is measured by looking at the codon frequencies in this reference set. CAI of any codon ranges between 0 and 1 such that for the synonymous codons coding each specific amino

acid the codon with CAI = 1 is the most advantageous one to use, in terms of translation efficiency, and the other codons with lower CAI values are less advantageous. In the first step a reference table of relative synonymous codon usage (RSCU) values is constructed:

$$RSCU_{ij} = \frac{n_{ij}}{\frac{1}{N_i}\sum_{j=1}^{N_i} n_{ij}} \qquad 1)$$

where the index *ij* indicates codon *j* in the amino acid *i*, and N_i is the number of synonymous codons, from 1 to 6, that the amino acid *i* is codded with. n_{ij} is the observed number of codon *j* coding the amino acid *i* in the genes that belong to the reference set, and the summation is over all the codons which code for the amino acid *i* in the reference set. RSCU value for a codon is simply the observed frequency of that codon divided by the frequency which we would expect from an impartial codon usage in each amino acid (Sharp et al., 1986). The relative adaptiveness of each codon is calculated by:

$$w_{ij} = \frac{RSCU_{ij}}{RSCU_{i\max}} = \frac{n_{ij}}{n_{i\max}} \qquad (2)$$

here the index '*imax'* indicates the codon which codes for the amino acid *i* and has the highest number compared to other synonymous codons. And it is obvious that since $RSCU_{ij}$ is proportional to Φ_{ij}, frequency of codon *j* coding for the amino acid *i*, we have:

$$w_{ij} = \frac{\phi_{ij}}{\phi_{i\max}} \qquad (3)$$

Finally the codon adaptation index of a specific gene can be calculated as:

$$CAI = \left(\prod_{k=1}^{l_g} w_k \right)^{1/l_g} \quad (4)$$

Where l_g is the number of codons, and w_k is the relative adaptivenes of the k^{th} codon in the gene sequence.

Using the same reference set, Ran & Higgs (2012) suggested a method for quantifying the strength of codon bias which improves the CAI measure. In this method they calculate logarithm of the ratio of the frequency of each codon in a reference set of highly expressed genes, which is assumed to be under translational selection, and their frequency averaged over the whole genome, where mutational bias is thought to be the dominating factor. For each codon *i*, the quantity:

$$\delta_i = \ln(\phi_i^H / \phi_i^0) \quad (5)$$

is defined, where ϕ_i^H and ϕ_i^0 are the frequencies of this codon in the high-expression set and the whole genome, respectively, measured as a fraction of the total number of codons for the corresponding amino acid. Codons with positive δ_i are preferred by translational selection relative to their synonymous codons. The δ measure for a gene is simply the average of δ_i for the codons in that gene. Genes with positive average δ, have codon frequencies that are similar to those in the high-expression reference set, and these are assumed to be under strong translational selection. The majority of the genes have a negative δ value, which means that their codon frequencies are more similar to the average genome frequencies than to the frequencies in the reference genes. The δ measure is

similar to the codon adaptation index (CAI), which also depends on ϕ_i^H, but δ specifically counts codons that increase in frequency in high expression genes as a result of selection, whereas CAI simply counts codons with high frequency in the reference set, which could be because of either mutation bias or selection (Ran & Higgs, 2012).

Dos Reis et al. (dos Reis et al., 2004), have introduced an index, tRNA Adaptation Index (tAI), for measuring how well, on average, the whole mRNA sequence can be translated. Since codon-anticodon pairing is not unique due to wobble interactions, more than one tRNA molecule might pair with each codon with different efficiency weights. Absolute adaptiveness of each codon is defined as follows:

$$W_i = \sum_{j=1}^{n_i} (1 - S_{ij}) tCGN_{ij} \tag{6}$$

Here n_i is the number of tRNA isoacceptors which could identify codon i. $tCGN_{ij}$ is the copy number of the j^{th} tRNA molecule which could pair with the codon i. S_{ij} is a parameter for considering the variation in coupling probabilities for different codon-anticodon combinations. All the efficiency weights, W_i, is divided by the maximal of all the 61 values to give the relative adaptiveness value, w_i, for each codon. Finally the tAI value of the gene g is calculated by geometrically averaging the relative adaptiveness of the codons in the sequence:

$$tAI_g = \left(\prod_{k=1}^{l_g} w_{i_{kg}} \right)^{1/l_g} \tag{7}$$

where the index i_{kg} indicates the codon i in the k_{th} position of the gene g. l_g is the length of

the gene *g*, in terms of codons.

The most challenging part of this index is finding the selective constraints on the codon-anticodon pairing, S_{ij}. A meaningful set of these values for each codon can be obtained by finding the values which maximizes the tAI in the highly expressed genes, since it is assumed that these gene are selected to show the highest adaptiveness possible.

1.4- Aims of this thesis:

In this thesis we focus on the 2nd and 3rd type of codon bias mentioned in section 1.2, codon bias among genes in a genome and different sites along the gene sequence, and try to test different scenarios for explaining the phenomena.

In order to see the causes behind the codon bias observed in the highly expressed genes, we focus on the proteome and transcriptome data for *E. coli* measured by Taniguchi et al., and the method introduced by Ran & Higgs for measuring the codon bias strength, and try to see among different measures of expression level, which one best explains the codon bias pattern we observe among the genes of *E. coli*. The data provided by Taniguchi et al. enable us to look at the different measures of expression, total number of produced protein molecules of each gene, total number of transcribed mRNA molecules of each gene, number of proteins produced per mRNA molecules of each gene, and also the protein production rate over each mRNA molecule at the same time. With the model introduced we could see at what level the codon bias has the most effect. There has been a huge debate on whether the protein production is elongation limited or

initiation limited and it has been shown that substitution of rare codons with frequent ones affects the elongation speed significantly (Sørensen et al., 1989), here we provide an analytic analysis to justify selection for frequent codons in spite of translation being initiation limited. We also test for context dependent mutation. We treat the synonymous codons as if the mutation rates are only affected by the third nucleotide position in codons or the second and third in order to test for the presence of context dependent mutation.

For the codon bias along the gene sequence, we focus on the appearance of rare codons in the beginning region of the genes. We specifically look at the relation between the folding free energy of the secondary structure in the beginning of the sequence and selection of the rare codons in this region. We hypothesize that the reduction in codon adaptation in this region is a side effect of selecting weak secondary structures in the translation initiation region of the genes. We have generated random synonymous sequences, sequences which are synonymous to the real genes but with different frequencies, to see how selection for weak folding in the beginning of ORF affects the codon usage in this region. For this matter we again focus on *E. coli* and by calculating the free energy of folding along the sequences we investigate the correlations between strength of secondary structure and the codon usage pattern observed in the genes.

Chapter 2

Causes of Codon Bias

2.1- Variations in codon bias strength among the genes in a genome (translational selection):

The causes for the pattern we observe in the codon usage bias we observe among the genes within a genome can lie between two extremes, mutational bias and natural selection (Hershberg & Petrov, 2008; Plotkin & Kudla, 2010). Even though there have been studies showing that mutational bias is a significant factor in shaping the codon bias (Kanaya et al, 2001; Knight et al., 2001 b; Chen et al., 2004) the fact that in almost all of the cases the preferred (most frequent) codon is the one with most abundant matching tRNA molecules, indicates that natural selection might play a role as well (Ikemura, 1985; Yamao et al., 1991; Kanaya et al., 2001, Higgs & Ran, 2008;). In fact the explanations which rely on natural selection can predict most of the patterns observed within a genome and the one which focuses on the mutational bias fits best with the codon usage variations between different species.

As an example, in *E.coli* by focusing on the two codon families we see that it's always the codon which benefits most from the tRNA pool (the C codon) that shows the highest frequency in the highly expressed genes and this preference is due to the fact that the tRNA molecules for these amino acids have a guanine in the wobble position which pairs best with the codon ending in C rather than the one ending in U (Sharp et al., 2005;

Higgs & Ran, 2008). Genes that use codons that can be coded by more numbers of tRNA molecules can be translated faster and/or more accurately, so they have an advantage over the ones that use the codons which don't have many tRNA molecules with appropriate anticodon to pair with. This advantage may be important for the genes coding for proteins whom the cell needs in large numbers in stages of rapid growth, resulting in the observed increase in strength of codon bias in highly expressed genes compared to the ones expressed in low levels, a fact also observed in other organisms such as *S.cerevisiae, C.elegans, Arabidopsis thaliana* and *D.melanogaster* (Ikemura, 1985; Yamao et al., 1991). The term "Translational Selection" refers to a process of selection on sequences for increasing the efficiency of their translation, rather than selection for functionality of the produced protein. Synonymous changes in gene sequences can affect the way a specific codon is translated, but does not affect the functionality of the resulted protein, and thus can affect the fitness of the organism in times of growth and reproduction (Higgs & Ran, 2008).

In the literature, there are different notions for translational efficiency on gene expression. Number of bound ribosomes per mRNA molecule (Ingolia et al., 2009); and number of proteins produced per mRNA (Tuller et al., 2010), that is, the ratio of protein abundance to mRNA level, are two famous measures introduced for this matter. The second definition is more relevant to issues of protein synthesis in each gene, whereas the former definition may be more relevant to ribosomal availability and overall cellular fitness. Weak correlation between these two notions of translational efficiency for endogenous genes indicates that the ribosomal density on a given mRNA molecule would

not show the amount of proteins produced from it (Plotkin & Kudla, 2010). It has also been reported that the average CAI of a gene in yeast, explains less than 3% of the variance in protein abundance per mRNA (Ingolia et al., 2009). Both of these observations support this school of thought that, for most endogenous genes, the initiation is the limiting factor for protein production (Bergmann & Lodish, 1979; Mathews et al., 2007). It has also been observed that the elongation speed of amino acid chain is significantly affected by insertion of preferred codons (the ones coded by more abundant tRNA molecules), into the mRNA sequence (Curran & Yarus, 1989; Sørensen et al., 1989). But it is not completely clear that increasing elongation speed in translation of one specific mRNA molecule can affect its total production rate significantly, since translation initiation rate, rather than elongation speed, might be the limiting factor in the process (Hershberg & Petrov, 2008; Plotkin & Kudla, 2010). However increasing elongation speed can reduce the time a ribosome spends on one mRNA and allow it to return to the pool of available ribosomes in the cell. This will increase the overall initiation and production rate of the genes in the cell, and therefore is beneficial overall. Simulations of protein production in Yeast show that increasing codon bias in a transgene could result in an increase in the pool of free ribosomes (Shah et al., 2013).

To see if the codon bias affects translation accuracy or speed, different studies have been conducted with results suggesting that the codon bias affects both parameters. The observation that in sites coding for more conserved amino acids, also show more bias in codon usage suggest that translation accuracy is affected by codon bias (Akashi, 1994; Stoletzki & Eyre-Walker, 2007). Akashi has found a preference for choosing the tRNA-

adapted codons at residues that are strongly conserved. Looking at *Drosophila* species, it was suggested that the sites which are under selection for conserving one specific amino acid, and thus selection for reducing the chance of mistranslation, also show codons which are most adapted with the tRNA pool. Using a broader group of species Drumond & Wilke have looked at the rate of evolution of different genes and correlation between the popper protein folding and parameters such as codon usage, gene expression etc. They have made the same observation as Akashi, and suggested that selection against mistranslation-induced misfolding is a sufficient factor for shaping the codon usage in highly expressed genes, in which an error in protein translation would be much more deleterious to the cell compared with lowly expressed ones.

2.2- Codon bias across one gene:

There are studies suggesting irregular codon usage in some specific organisms or special sites in the genes, but recent studies suggest other patterns of codon usage across a gene which is thought to be shared between diverse species (Plotkin & Kudla, 2010).

Bulmer and Eyre-Walker, motivated by the work of Burns & Beacham, were among the first to derive a translation efficiency profile of the codons in the genes sequences (Bulmer, 1988; Eyre-Walker & Bulmer, 1993). Their findings clearly show a significant reduction in the CAI value of the first 20-30 codons of the genes, compared to the rest of the sequence, and this reduction becomes more significant as the average CAI of the genes increases (Figure 2.1).

There are two competing theories for explaining this phenomenon. One regards this bias as a mechanism for slowing elongation rate in the beginning of the translation of peptide chains in order to regulate the movement of the ribosomes along the mRNA (Tuller et al., 2010), and the other one treats the observed translation efficiency profile as a side effect of selecting for weak folding in the translation initiation region of mRNA sequence (Eyre-Walker & Bulmer, 1993; Bentele et al., 2013).

Different studies have showed the importance of mRNA secondary structure in the ribosomal binding site on the initiation of the protein translation and generally on the protein production rate (Bentele et al., 2013; Kudla et al., 2009; de Smit & Van Duin, 1990). Strong secondary structure near the initiation region of the mRNA sequence could affect protein production in two ways: First, strong local mRNA secondary structure would have a negative impact on the ribosomal binding rate. Second, if the start-codon is captured in the middle of the folded region, the ribosome would be unable to recognize it (Gu et al., 2010). Gu et al., claim that the latter affects the process of translation initiation more significantly than the other.

Gu et al., and more recently, Bentele et al., have measured the folding energy in different parts of the gene sequences in diverse species and have detected a selection for weak secondary structure in the translation initiation region of mRNA sequences (Figure 2.2). The reduction in folding strength of the mRNA in the beginning of the sequence can be well predicted by the total GC content of the genome. As the GC content increases, the suppression of the secondary structure in translation initiation region increases (Gu et al.,

2010). Besides a strong correlation between the suppression of mRNA secondary structure near the translation initiation region and the deviation in codon usage in the same region compared to the rest of the sequence has been found. There is also a pattern of reduction in total GC content and GC3 (GC content in the third position of the codons), in the beginning of the genes, which would be expected since guanine and cytosine would create a much stronger bond compared to adenine and uracil and therefore cause a stronger folding. And since in GC rich organisms, such as *E. coli*, the abundant codons tend to rich in GC and a reduction in GC content in the beginning of the ORF, in order to reduce the folding energy of the secondary structure, will result in using AU rich codons which are rare (Bentele et al., 2013).

Tuller et al. findings on the efficiency profile of codons in different species using tRNA adaptation index, tAI, also show a clear selection for choosing inefficient codons for the first 30-50 codons, Figure 2.3. They term this region "ramp", and the statistical tests clearly show that the ramp is selected for. But they provide a different explanation for the existence of this phenomenon. According to their argument, the ramp is a mechanism to control the movement of the ribosomes along the mRNA sequence.

Both experimental measures and simulations show that insertion of a segment of rare codons in the middle of a gene could affect the translation efficiency of the gene significantly, since queuing of ribosomes behind this region can occur and thus would cause a bottleneck in protein translation (Shaw et al., 2004; Mitarai et al., 2008). Introducing a region of slow codons in the beginning of the sequence will cause spacing

between ribosomes along the sequence and therefore decrease the chance of jamming of ribosomes when encountering the bottlenecks during protein translation. This would be beneficial since one factor involved in the cost of translation of proteins would be the total time a ribosome spends on each mRNA molecule, and reducing the chance of collisions would save the ribosomes from wasting time on the sequence. Besides, the ramp may as well increase the sensitivity to the abundance of tRNA molecules loaded with amino acids at early stages of translation process and thus provide a simple way of terminating the translation process in the beginning in the case of insufficient level of raw materials. A negative correlations between the total number of transcribed mRNA molecules and number of ribosomes bound per mRNA, with the length and depth of the ramp has also been detected, which would support this explanation for the existence of the ramp since the jamming of ribosomes would be more dramatic for genes which have higher mRNA levels and higher number of ribosomes per mRNA.

2.3- Figures & Tables

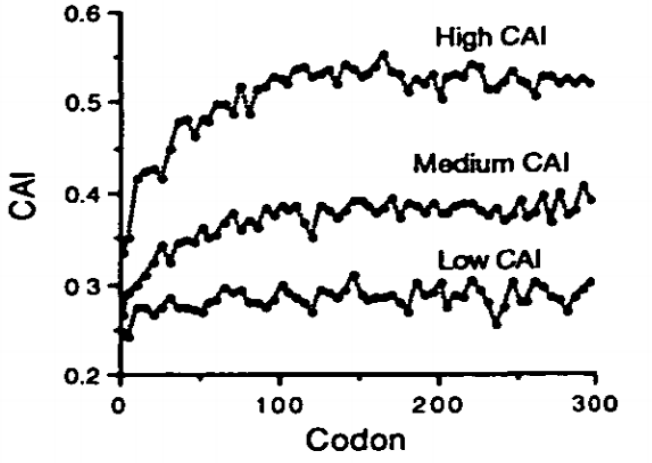

Figure 2.1: CAI profile of E. coli genes for each codon position, divided into three groups according the average CAI value of the sequence. *(Eyre-Walker & Bulmer, 1993)*

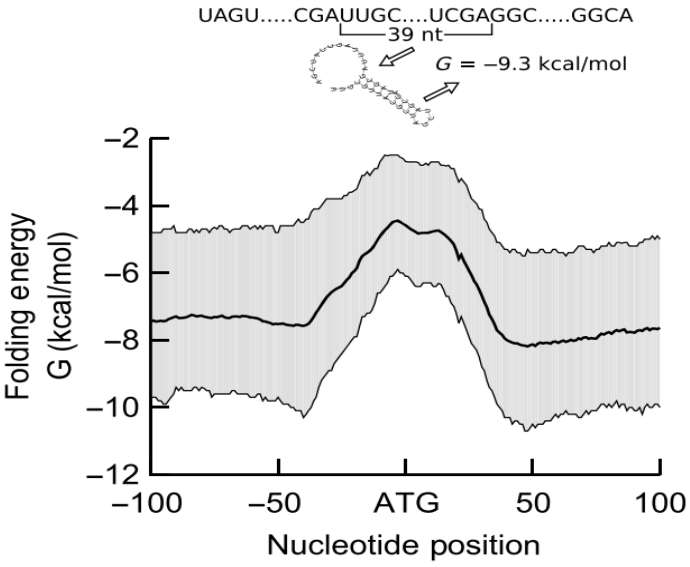

Figure 2.2: the profile of secondary structure folding energy in mRNA sequence of E. coli. The average folding energy shown in solid line with an interquartile range in grey. *(Bentele et al., 2013)*

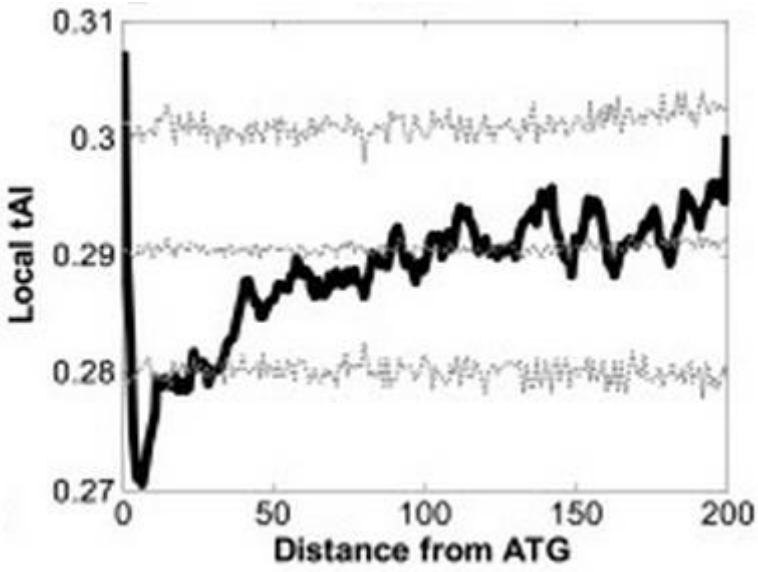

Figure 2.3: A region of codons with low tRNA adaptation index (tAI) at the beginning of *E. coli* gene sequences. *(Tuller et al., 2010)*

Chapter 3

The Relationship Between the Strength of Codon Bias in Gene Sequences and the Expression Level of the Corresponding Proteins and mRNAs

3.1- Introduction:

Looking at the codon frequencies in different genes in *E. coli* we would observe a clear bias in choosing the synonymous codons, and this bias becomes more significant in the highly expressed genes. There are different theories which try to explain this phenomenon; some would refer to the translational selection as the dominant force shaping this bias and some focus on the mutational bias. Here by using the data on the proteome and transcriptome of *E. coli* and using a population genetics model we try to investigate the relation between codon bias in a gene and different measures of expression level.

Taniguchi et al. 2010, have reported single-cell global profiling of both mRNAs and proteins using a yellow fluorescent protein (YFP) fusion library for *E. coli*, and their data has enabled us to look at the relation between different measures of expression level and codon usage bias in the gene in *E. coli*. In this we can see whether increasing the strength of codon bias in a gene affects parameters related to its own protein production or the overall protein production and fitness of the cell.

Several studies suggest that the protein production of each gene is initiation limited and substitution of rare codons with preferred ones increases the elongation speed, but may not necessarily increase the overall protein production of the gene itself (chapter 2, section 2.1), therefore selection for stronger codon bias in highly expressed genes may not seem intuitive. Here we also try to suggest an analytic model for protein production which allows for selection of preferred codons in spite of translation being initiation limited.

Our model also enables us to test for the existence of context dependent mutation. Signatures of context dependent mutation has been observed in many organisms (Jia & Higgs, 2008; Shioiri & Takahata, 2001; Fedorov et al., 2002), suggesting that mutation rates between the 4 nucleotides in the gene sequences are affected by the neighboring sites. In this model we are able to see whether the second nucleotide position in the synonymous codons could affect the mutation rates between the codons coding for an amino acid.

3.2- Expression Measures in *E. coli*:

In this study we have aimed to analyze dependence of different features of codon bias in *E.coli* on gene expression. We have used the results given by Taniguchi et al., in which they have measured average protein production rate for 1018 genes, and for 585 genes out of the 1018 genes they have measured average mRNA levels and also mRNA lifetimes in a cell cycle.

Taniguchi et al., used the following mathematical model to describe the concentrations of proteins and mRNAs in the cell. Here we use this model to generate four

different hypotheses as to how the strength of selection of codon bias might vary among genes.

Let p and m be the mean number of copies per cell of a specific protein and its mRNA. These satisfy the differential equations

$$\frac{dm}{dt} = k_1 - \gamma_1 m\,, \qquad \frac{dp}{dt} = k_2 m - \gamma_2 p\,,$$

where k_1 is the transcription rate, k_2 is the translation rate over each mRNA, and γ_1 and γ_2 are the breakdown/dilution rates of the mRNA and protein. It is assumed that the major factor leading to dilution of proteins is the growth and division of the cell, so that $\gamma_2 = 1/T_{cell}$ for all proteins, where T_{cell} is the cell division time. The mRNA breakdown rate is $\gamma_1 = 1/T$, where T, the mRNA lifetime, is different for each mRNA and can be substantially shorter than T_{cell}. In steady state, we have

$$m = \frac{k_1}{\gamma_1} = k_1 T\,, \qquad p = \frac{k_2 m}{\gamma_2} = m k_2 T_{cell}\,.$$

It is also useful to define M and P, the mean number of mRNAs and proteins produced per cell cycle. It follows that

$$M = k_1 T_{cell} = \frac{m T_{cell}}{T}\,, \qquad P = m k_2 T_{cell} = p\,.$$

Taniguchi *et al.* fit the distribution of fluorescence intensity between cells using two parameters a and b, where a is the number of mRNAs produced per cell cycle (which

iscalled *M*), and *b* is the mean number of proteins produced from one mRNA. It follows that

$$b = k_2 T\,, \qquad P = Mb\,.$$

In the experiment, *P*, *M*, *b* and *T* are all measured for many different genes in *E. coli*. Here, we test four hypotheses about the way the strength of translational selection should depend on these quantities.

I. $S \sim P$.

II. $S \sim M$.

III. $S \sim b = P/M$

IV. $S \sim k_2 = P/(MT)$

In these hypotheses, *S* is the strength of selection that appears in the mutation/selection/drift theory of codon usage bias. As the total effort expended on synthesizing a protein is *P*, it seems clear that *S* should depend on *P* (Hypothesis I). As *P* and *M* are correlated [(Taniguchi et al., 2010) & Figure 3.1], it also seems reasonable that *S* should depend on *M* (Hypothesis II). If codon usage bias arises as a result of selection for translational efficiency, then it also seems reasonable that genes with a higher proportion of fast codons should produce more proteins per mRNA; hence, *S* should depend on *b* (Hypothesis III). Finally, we would expect that codon bias should influence translational rate per mRNA, k_2 (Hypothesis IV). We note that all four quantities are correlated, so we should not be surprised if all four hypotheses are true to some extent. Therefore, we

consider quantitative predictions of codon frequency data using models based on the four hypotheses in order to determine which factors are most relevant in determining codon bias.

There is an important caveat regarding Hypothesis IV. In the simple dynamical theory above, translation is treated as a single process with a rate k_2. This is a gross oversimplification. Translation involves both initiation and elongation. Initiation (*i.e.* the binding of a ribosome to the 5' end of a mRNA and moving to the first codon) is likely to vary in rate between different mRNAs in ways that are not directly related to codon bias. Codon bias should be directly related to the elongation rate; however, the data that we use from Taniguchi *et al.* do not measure elongation rate, so we cannot use these data to test a hypothesis that *S* is dependent on elongation rate. Later in this chapter, we consider a more detailed theory which distinguishes between initiation and elongation. We wish to emphasize that selection should still occur on codon usage even when translation is initiation-limited. At this point we proceed to the data analysis using the hypotheses that are testable from the data of Taniguchi et al.

3.3- Correlation of Expression Level with Codon Bias

We have downloaded the genome of *E.coli* from NCBI database and calculated codon frequencies of each gene for further analysis of dependence of codon bias on gene expression. As a measure of the strength of codon bias, the average δ was measured for each gene, as suggested by Ran & Higgs (2012), and discussed in Chapter 1.

3.4- Correlation between different measures of expression level

The data extracted from Taniguchi *et al.* show clear positive correlation between mean protein level and mean mRNA level, mean protein per mRNA and also mean protein per mRNA per unit time, Figure 3.1. This fact could indicate that in order to increase the mean protein level, the cell tries to both increase transcription and translation speed. As it can be extracted from Figure 3.1, since P/M depends on T itself, dependence of the protein level in each gene can be traced back to three independent parameters: number of transcribed mRNA molecules (M), protein translate rate over each mRNA molecule (P/(M×T)) and the lifetime of mRNA molecules (T).

3.5- Codon bias is correlated with P, M and other expression level measures

Plotting the measure of δ for strength of codon bias, averaged over each gene against protein level shows the strong correlation of codon bias and protein production level, Figure 3.2. This plot shows that even for the genes with very low production level there is a selection for choosing codons with high δ which are preferred in the highly expressed genes. We also observe that strength of codon bias and other expression factors show less correlation; this may support the idea that increase in protein production rate over each mRNA and increasing elongation rate of proteins, something that increasing strength of codon bias could result to, are not strongly correlated (Hershberg & Petrov, 2008; Plotkin & Kudla, 2010).

3.6- Population genetics theory for codon frequencies

The population genetics theory for the way codon frequencies should depend on mutation, selection and drift goes back to Bulmer (Bulmer, 1991) and has been used by several authors (Ran & Higgs, 2012; Shah & Gilchrist, 2011; Trotta, 2013). The expected frequency of codon *i* in gene sequence *g* can be written as

$$\phi_{ig} = \frac{\mu_i \exp(S_{ig})}{\sum_j \mu_j \exp(S_{jg})} \qquad 8)$$

where μ_i is the mutation rate to codon *i* from its synonymous codons (which is assumed to be independent of the gene), and S_{ig} is the scaled selection strength acting on codon *i* in gene *g* (which depends on *g* because different genes have different expression levels). The sum in equation 8 is over all the codons *j* that are synonymous with *i*.

The scaled selection strength can be written as $S_{ig} = 2N_e s_{ig}$, where N_e is the effective population size, and s_{ig} is the selection coefficient in the fitness. However, N_e cannot be determined from codon frequency data, so we deal directly with S_{ig}. If $S_{ig} << 1$ for all codons in a family, then the codon frequencies depend on mutation rates only:

$$\phi_{ig} = \frac{\mu_i}{\sum_j \mu_j} \qquad 9)$$

If selection is strong, then the codon frequency ϕ_{ig} tends to 1, for the codon that has the highest S_{ig} in the codon family.

3.7- Testing hypotheses for mutation and selection

The aim of this paper is to compare several alternative hypotheses for the way the mutation and selection parameters in this theory should depend on codons and sequences. If n_{ig} is the observed number of occurrences of codon i in gene g, then the log likelihood, L, of the set of genes is

$$\ln(L) = \sum_g \sum_i n_{ig} \ln(\phi_{ig}) \qquad 10)$$

The free parameters in the theoretical frequencies, ϕ_{ig}, are chosen to maximize the likelihood. To select between models with different numbers of parameters, we use Akaike's information criterion, AIC = 2(-ln L + K), where K is the number of free parameters to be estimated from the data. The model with the minimum AIC is to be preferred (Akaike, 1998). AIC selects models with high likelihood but penalizes models with unnecessarily large numbers of parameters.

If no assumptions are made about the mutation rates, then there is a different μ_i parameter for each codon. However, the codon frequencies depend only on the relative rates of forward and reverse mutations, not on the absolute mutation rates. Therefore it is possible to set $\sum_j \mu_j = 1$ for every codon family. For a family of n synonymous codons, there are n-1 independent μ_i parameters. In the standard genetic code, there are 3 families with 6 codons, 5 families with 4 codons, 1 family with 3 codons, and 9 families with 2

codons. This gives 41 μ parameters. We exclude stop codons and codons for single-codon amino acids, Met and Trp. We call this the general-μ assumption.

A first hypothesis for selection is that it should be linearly proportional to protein level, $S_{ig} = k_i P_g$, where P_g is the measured protein level, and there is a selection parameter k_i for each codon that is to be estimated from the codon data. Codon frequencies depend on relative selection values; hence we chose to set $k_i = 0$ for the most preferred codon for each amino acid in the high expression genes. There are n-1 independent k parameters in a family of n codons, and 41 independent k parameters in total. We call this the general-k assumption. Our initial model, with general-μ and general-k assumptions and selection linear in P, has K=82 free parameters. We compared alternative models with fewer μ or k parameters relative to this model using AIC. The alternatives are defined in the results section.

As there is considerable fluctuation in codon numbers, it is sometimes useful to smooth the data by binning genes according to protein level, and averaging over genes in each bin. Genes were ranked by P_g and divided into 14 bins with equal numbers of genes in each bin (the choice of number of bins does not affect the final results qualitatively, and 14 was picked so that the difference between last bin and the rest, in terms of number of genes included in each bin, would be minimized). P_m is the mean value of P_g for genes in bin m and n_{im} is the number of occurrences of codon i in genes in bin m. If selection is linear in P_m, then $S_{im} = k_i P_m$. The formulae for the codon frequencies ϕ_{im} and ln L are equivalent to equations 1 and 3 with the bin index m replacing the gene index g.

By binning the genes and treating the parameters such that the theoretical frequencies would initially go through the observed codon frequencies in two of bins, one picked from high expression regime and one from low expression regime, we were able to make an initial guess for the parameters involved in each model. The parameters were then varied such that the likelihood of the model reaches its maximum.

We also considered models where selection depends on the mRNA level, *M*, instead of the protein level *P*, or on some combination of *P* and *M*. As these quantities are experimentally measured, changing these assumptions about selection does not change the number of parameters.

3.8- The variation of individual codon frequencies with protein level

As Figure 3.2 indicates, codon bias depends more strongly on *P* than other measures of expression level, we first tested evolutionary models that assume translational selection strength is dependent on *P*. Genes were binned according to their total protein level, and we supposed that the selection strength S_{im} is linearly dependent on P_m, the standard model. The standard model has general μ and general *k* parameters, and selection is linear in protein concentration, - $S_{im} = k_i P_m$. The calculated values extracted from the model shows very good agreement with what we actually observe in the real genes.

Next we tested for different functions for the relation between fitness and protein level: $S_{im} = k_i \, ln(P_m)$, $S_{im} = k_i (P_m)^{\alpha}$ (with α being a free parameter) &

$S_{im} = k_i \frac{P_i}{1 + P_i / P_{sat}}$ (with P_{sat} as a free parameter) . Among all these functions, the latter shows the least AIC value (when $P_{sat} \approx 1.9 \times P_{average}$). This shows that the linear dependence of selection strength on protein level is only valid in lowly expressed genes, and as the gene expression is comparable with the average protein level, the dependence of fitness on protein level starts to deviate from the linear function and reaches a final limit for large protein levels. The table of AIC values is given in Table 3.1.

One thing that could affect the codon frequency significantly is the context-dependent mutation, which indicates that the mutation between one nucleotide and any other nucleotide is partly controlled by the neighboring sites. This model allows us to test context-dependent mutation as well. We observe that in Tyr, His, Asn & Asp the frequency of codons having the same nucleotide in the second and third position (AU and AC), in the genes with low expression level, which is assumed to be mainly governed by mutation, differ significantly.

For testing the existence of context dependent mutation we have treated mutation of codons in two ways. First we tested whether by only considering the third position as the main parameter determining mutation rates, we can regenerate the bias in codon frequencies we observe. We looked at codon frequencies in the very lowly expressed genes, which are assumed to be governed mainly by mutational bias, and treated all the codons with same nucleotide in the third position alike. All the codons ending with the same nucleotide, would show the same mutation rate. In this manner the 41 independent parameters for μ, would decrease to 3. By doing this we observe that the likelihood

decreases by a significant amount and this method shows a much higher AIC value, indicating that the reduction of the number of parameters destroys significant amount of information. Next we tried a different approach. We tried to see if considering the second and third position as the two main sites controlling the mutation rate would give an acceptable result. This time we treated all the codons in each column of the genetic code, which share same nucleotides in the second and third positions, alike. In this way the number of independent μ parameters would be 12, 3 independent parameters in each column, instead of 41. This time as well the final likelihood is much less than the most general case and a higher AIC value. This shows that for each codon the mutation rate is context dependent and synonymous mutation rates observed in the genes cannot be explained by just considering the third position or second and third position. The values for μ parameters of the codons in *E. coli* are presented in Figure 3. 4, for two and four codon families separately.

We also tested whether in the U+C codon families (Phe, Tyr, His, Asn, Asp & Cys), the preferred codons could all have the same selection coefficient or the ones which have the same number of tRNA genes, UAC/GAC and CAC/UGC, with perfect codon-anticodon pairing can have the same value as their selection coefficient. But we also observe that these two models would again result in an increase in AIC (Table 3.1).

Next we tried to see how does considering other measures of expression for the translation selection strength affects the results. We tested different functions for the translation selection and compared the AIC value of each model with the model where the

selection has a linear dependence on P. The AIC values are given in Table 3.2. As it can be extracted from the data, the translational selection in mainly dependent on the total protein level of the gene, however considering the number of transcribed mRNA molecules of each gene could result in an improvement to the fit.

The result of fitting the data on *E.coli* with the model with lowest AIC is shown in Figure 3.3, for Phenylalanine and Valine. As it can be seen the model can fit through the data points pretty well and can produce the pattern we observe in *E.coli* genome.

3.9- Testing the model for Yeast:

In order to see whether the proposed model works in other organisms as well we have looked at the genome of yeast (downloaded from NCBI) and by using the data measured by Ghaemmaghami et.al, protein level of more than 3800 genes, we could extract the relation between codon bias and gene protein level in yeast, Figure 3. *5*. The genes used were binned into 14 bins and for each bin the frequency of each codon and the average protein level was calculated and finally was fitted by the model, the results for Phenylalanine and Valine are provided in *Figure 3.6* as an example. Furthermore as it can be seen in Table 3.3, putting restriction on mutation rate causes a significant increase in AIC, an indication for existence of context dependent mutation in yeast. Again we see that the function with a saturating function of protein level fits best with the data.

3.10- Effects of initiation and elongation on protein production rate

There has been some debate about whether translation is limited by initiation or elongation. The term 'initiation-limited' has been used rather imprecisely in the literature, which has contributed to some confusion. A useful way to define what is meant by 'initiation-limited' is to consider the TASEP model (totally asymmetric simple exclusion process), which is often used as a model of translation of a single mRNA by multiple ribosomes simultaneously. In models of this type, it is assumed that ribosomes bind at the 5' end of the sequence at a rate α, provided the initial region of the sequence is not already occupied by a previous ribosome. A ribosome moves forward one codon at a rate v, provided its progress is not blocked by another ribosome immediately in front of it. In the simplest versions of the model, this rate is the same for every codon, but in more realistic versions, different rates can be assigned to codons of different type (Zia et al., 2011; Chou & Lakatos, 2004; Shaw et al., 2004).

The rate of protein production from one mRNA is equal to the current, J, of ribosomes moving along the mRNA. In a state of steady translation, J is the same at all points along the mRNA. For every ribosome that initiates, another one terminates at the end of the sequence. Thus the protein production rate is equal to α times the probability that the initiation region is not blocked by a previous ribosome. If α is small compared to v, the ribosomes are well separated along the mRNA and the initiation region is always free of previous ribosomes. This means that $J = \alpha$ in this limit, which we will refer to as strictly initiation-limited. For somewhat larger α, the ribosomes are fairly well separated, but there

is a chance that a previous ribosome blocks the initiation region, so that the successful initiation rate is reduced, and J is slightly less than α. J is a function of both α and v in this regime, which is known as the low density phase. If α is increased further, there is a high probability that the initiation region is blocked. In this case there is a maximum current that is proportional to v and independent of α (Shaw et al., 2004). In this maximum current phase, protein production is strictly elongation-limited and is independent of the attempted initiation rate. The situation is more complex if there is significant variation in rates between fast and slow codons within a sequence. In some cases, the current can be controlled by bottlenecks of a few particularly slow codons, rather than by the average elongation rate v.

The regimes limited by elongation and by bottlenecks are interesting as dynamical phenomena and have been widely studied in simulations (Greulich & Schadschneider, 2008; Dong et al., 2007; Shaw et al., 2003). However, it is the elongation limited regime that seems most relevant to understanding the evolution of codon usage bias in highly expressed genes. Rapidly multiplying microorganisms are under selection for increasing overall protein production rate (Ran & Higgs, 2012). This depends on translating lots of different mRNAs simultaneously with a finite number of ribosomes. Here we wish to give a simple analytical theory that applies to simultaneous translation of many different mRNAs in the limited translation is strictly initiation-limited.

We suppose that the total number of ribosomes in a cell at a given time is N_{tot}, and that N_{free} of these ribosomes are free to initiate translation. The number of ribosomes already bound to mRNAs and engaged in translation is $N_{bound} = N_{tot} - N_{free}$. The mean

number of mRNAs for gene g in the cell is m_g and the total number of mRNAs is $m_{tot} = \sum_g m_g$. We suppose that ribosomes are well separated along mRNAs. The rate of initiation of ribosome on an mRNA for gene g is $\alpha_g = r_g N_{free}$, where r_g is an initiation rate constant that depends on the gene, and it is assumed that initiation increases linearly with the number of free ribosomes. The time taken for one ribosome to translate an mRNA is t_g. For widely spaced ribosomes, this is just the sum of the times taken for each codon. The mean time per codon is t_g/L_g, where $L_g$is the length (in codons) of the gene. The mean elongation rate is $v_g = L_g/t_g$.

The mean time between initiation events on the same mRNA is $1/\alpha_g$, and the time spent by one ribosome on the mRNA is t_g; therefore the mean number of ribosomes bound to the mRNA is

$$n_g = \alpha_g t_g = N_{free} r_g t_g \qquad 11)$$

and the mean separation, d_g, between ribosomes is the number of codons moved by one ribosome in a time $1/\alpha_g$:

$$d_g = \frac{v_g}{\alpha_g} = \frac{L_g}{t_g r_g N_{free}} = \frac{L_g}{n_g} \qquad 12)$$

Each ribosome covers a length of l codons, and it is estimated that l is approximate 11 in *E. coli*. Hence, the condition that the ribosomes are widely separated, and do not interfere with each other is that $d_g >> l$, or equivalently, $\alpha_g << v_g/l$. This is what is meant by

saying that the process is initiation-limited. It does not imply that the time between initiations is short compared to the time spent on the mRNA. In fact, if $n_g > 1$, $t_g > 1/\alpha_g$.

If we are in the initiation-limited regime, the number of free ribosomes can be obtained in the following way.

$$N_{tot} = N_{free} + N_{bound} = N_{free} + \sum_{g'} m_{g'} n_{g'} = N_{free}(1 + \sum_{g'} m_{g'} r_{g'} t_{g'}) \quad (13)$$

$$N_{free} = \frac{N_{tot}}{1 + \sum_{g'} m_{g'} r_{g'} t_{g'}} \quad (14)$$

We use g' as an index summing over all genes. For any one particular gene, g, we can obtain the number of ribosomes per mRNA, n_g, and the rate of protein production per mRNA, J_g:

$$n_g = \frac{N_{tot} r_g t_g}{1 + \sum_{g'} m_{g'} r_{g'} t_{g'}} \quad (15)$$

$$J_g = \frac{N_{tot} r_g}{1 + \sum_{g'} m_{g'} r_{g'} t_{g'}} \quad (16)$$

The number of copies of the protein produced per cell cycle is:

$$P_g = m_g J_g T_{cell} = \frac{N_{tot} T_{cell} m_g r_g}{1 + \sum_{g'} m_{g'} r_{g'} t_{g'}} \qquad 17)$$

and the total number of proteins produced per cell cycle is

$$P_{tot} = \frac{N_{tot} T_{cell} \sum_{g'} m_{g'} r_{g'}}{1 + \sum_{g'} m_{g'} r_{g'} t_{g'}} \qquad 18)$$

Consider a synonymous mutation that reduces the time spent by the ribosome on this codon by an amount δt. If the mutation occurs in a specific gene g, the time for translation of this gene is reduced from t_g to t_g-δt. The denominator in equations 7-11 depend on the times $t_{g'}$ for all the genes. When the mutation occurs, it will increase the production rate P_g of the gene g and it will also increase the rate of production $P_{g'}$ of all the other proteins by the same factor. Let ΔP_{tot}^{g} be the amount by which the total protein production rate is increased due to a mutation occurring in gene g.

$$P_{tot} + \Delta P_{tot}^{g} = \frac{N_{tot} T_{cell} \sum_{g'} m_{g'} r_{g'}}{1 + \sum_{g'} m_{g'} r_{g'} t_{g'} - m_g r_g \delta t} \qquad 19)$$

Assuming that δt is small, and expanding to first order in δt, we obtain

$$\Delta P_{tot}^{g} = P_{tot} \frac{m_g r_g \delta t}{1 + \sum_{g'} m_{g'} r_{g'} t_{g'}} = \frac{P_{tot}}{N_{tot} T_{cell}} P_g \delta t \qquad 20)$$

The key point is that if the same mutation with the same δt occurs in any gene, its effect on the protein production rate is proportional to P_g. Thus, if selection is acting to increase the overall growth rate of the cell, and if the time spent on protein production is a significant proportion of the total time required for cell division (which seems likely), then the fitness change due to the mutation should be proportional to the protein production rate P_g of the gene in which it occurs. This motivates Hypothesis I: $S \sim P$. It also explains why selection acts on codon usage even when translation is initiation-limited.

3.11- Discussion and Conclusion:

The dependence of codon bias strength and the expression level of the genes have been suggested by several scientists (Ikemura, 1985; Bulmer, 1991; Shah & Gilchrist, 2011; Ran & Higgs, 2012), here we try to further investigate this phenomenon and see what measure of expression level would explain the differences in codon bias strength among different genes in a genome.

Using the data measured by Taniguchi et al., and the population genetics model we could be able to compare different models considering different expression measures as the dominant force in translational selection. The results indicate that total protein level of the genes explains the codon bias pattern observed in the real genes better than other parameters, M, P/M & P/ (MT). We first proposed a linear dependence of translational selection strength on the total protein level, $S \sim P$. But it was realized that this is true only for lowly expressed genes and as the protein level grows, the translational selection saturates and reaches a final limit. Considering the number of transcribed mRNA

molecules in each gene, beside the total protein level improved the model by a small amount. We have introduced a model for protein production which assumes that the protein translation is limited by initiation rate, and not the elongation speed. The analytic analysis show that the effect of a synonymous substitution in the codons of the gene sequences, which would cause a faster elongation speed, would increase the total number of proteins produced in the cell by a factor which is proportional to the total protein level of that specific gene. This would justify the higher correlation observed between codon bias strength and the total protein production of the genes compared to other measures of expression.

As Table 3.1 shows, in highly expressed genes the dependence of strength of selection on the protein level saturates, it reaches a final value of P_{sat} for very large protein levels. One reason for this phenomenon could be that reaching the highest possible level of frequent codons would not be optimal, but rather a specific fraction of frequent/rare codon frequencies would result in the highest fitness of the cell. It has been reported (Kolmsee & Hengge, 2011), *rpoS* sigma factors of *E. coli* contain large numbers of rare codons and substitution of these rare codons with the frequent codons results in a reduction of level of mRNA transcription and protein production. By reducing the speed of the ribosomes, rare codons can regulated spacing of the ribosomes on mRNA molecules and thus prevent the transcript from the ribonucleolytic attacks.

By checking whether the model is over parametrized, we were also able to test the existence of context dependent mutation. The results show that the mutation rates

calculated can only be explained if we consider all three nucleotide positions of the codons, which suggests the existence of context dependent mutation. This phenomenon is also observed in yeast. We could also see whether reducing the number of selection coefficients, k_i, would affect the model. We see that even setting selection coefficients of the codons with the same number of tRNA genes with proper anit-codon pair, in U+C two codon families, would cause a significant loss in the information suggesting that the translation of each codon depends on the details of the codon-anticodon interactions or the number of tRNA molecules is not perfectly proportional to the number tRNA genes and thus is not sufficient to explain the translation efficiency of each codon.

The mutation rates and selection coefficients of each codon, along with other frequencies and number of tRNA genes are given in Table 3.4.

3.6- Figures & Tables:

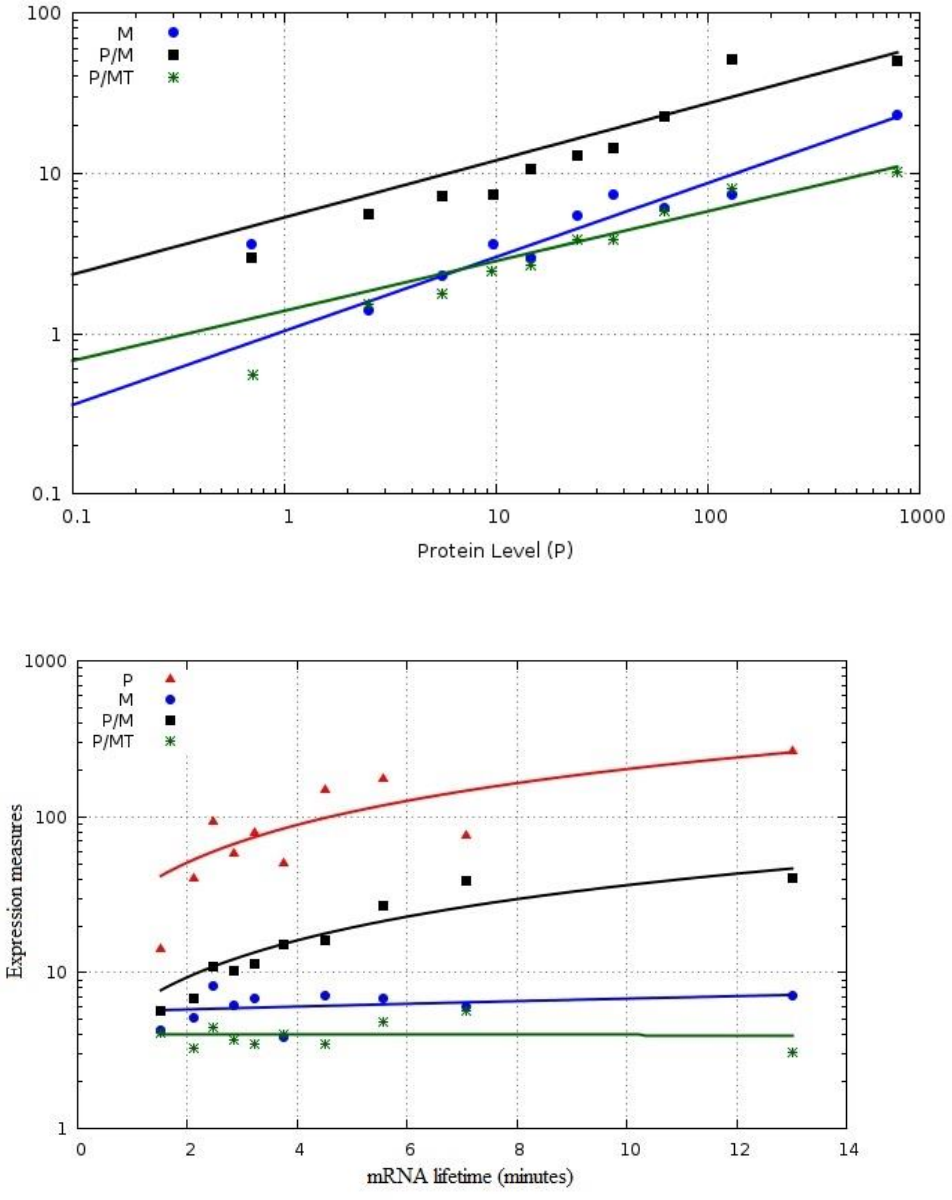

Figure 3.1: Correlation between different expression measures level, M, P/M, P/(M×T) [where T is the mRNA lifetime], and total protein level (top); and the correlation between P, M, P/M, P/(M×T) and mRNA lifetime, bottom. (data from Taniguchi et al., 2010)

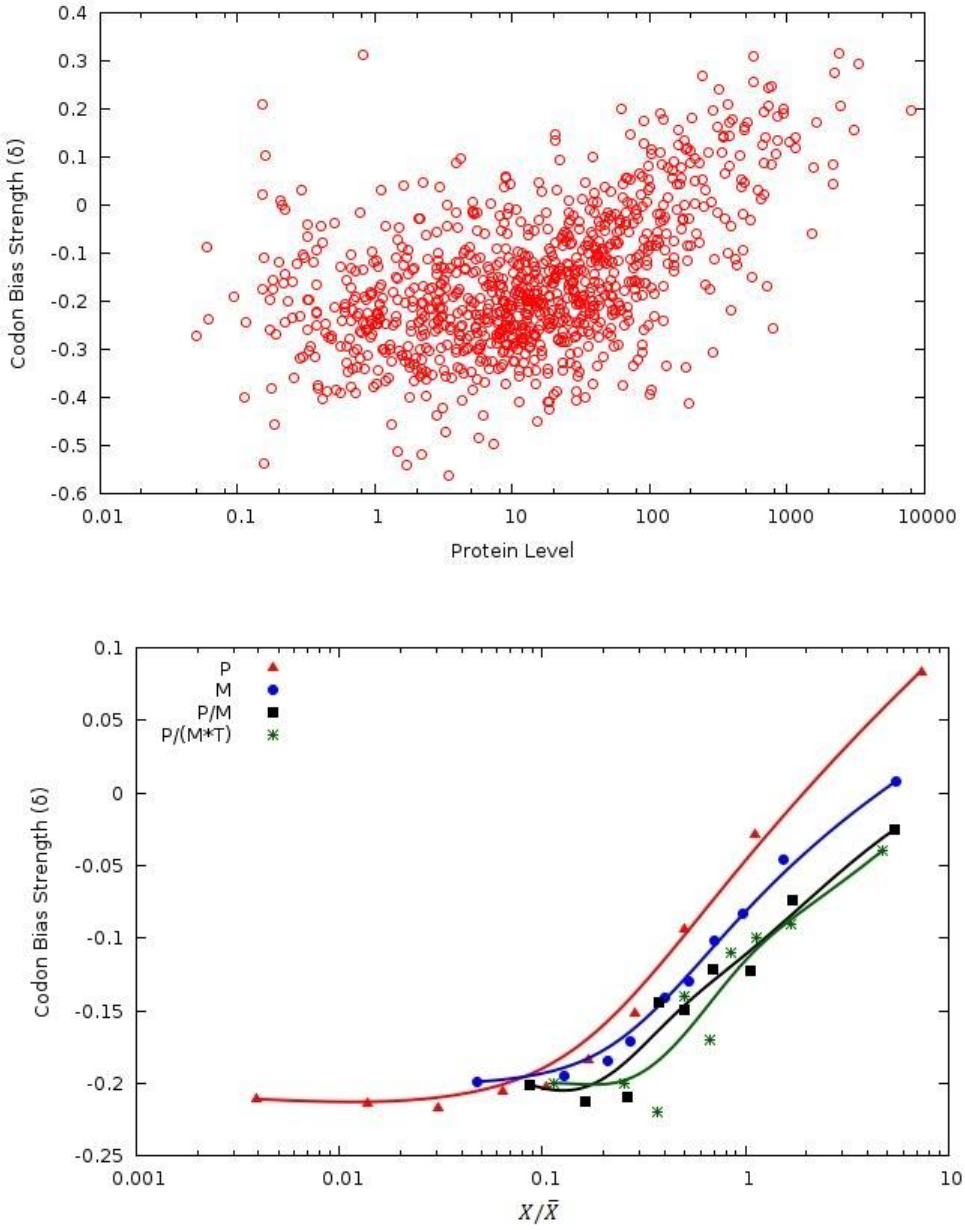

Figure 3.2: Correlation between the codon bias strength (δ) and protein level of the genes, top, and other expression measures, bottom. For the plot in the bottom, in order to be able to compare the different expression measures, the parameters (X) were divided by their average ($\bar{X}$) values so that they would have the same scale.

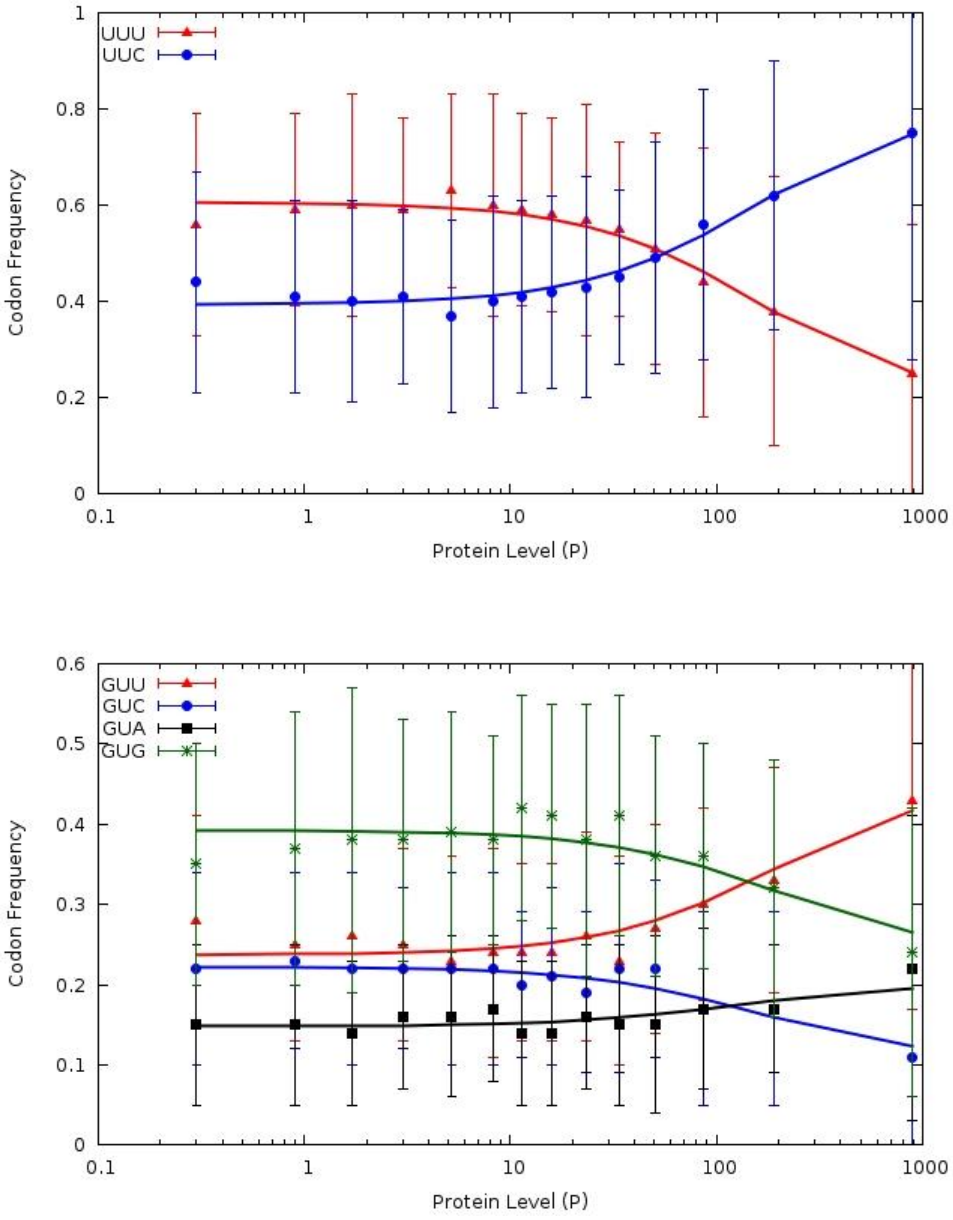

Figure 3.3*:* Codon frequency pattern for Phenylalanine, top, and Valine, bottom. Markers show the observed frequencies in the genome, whereas the solid lines show the values from the model. The error bars shows one standard deviation in frequency of each codon in each bin.

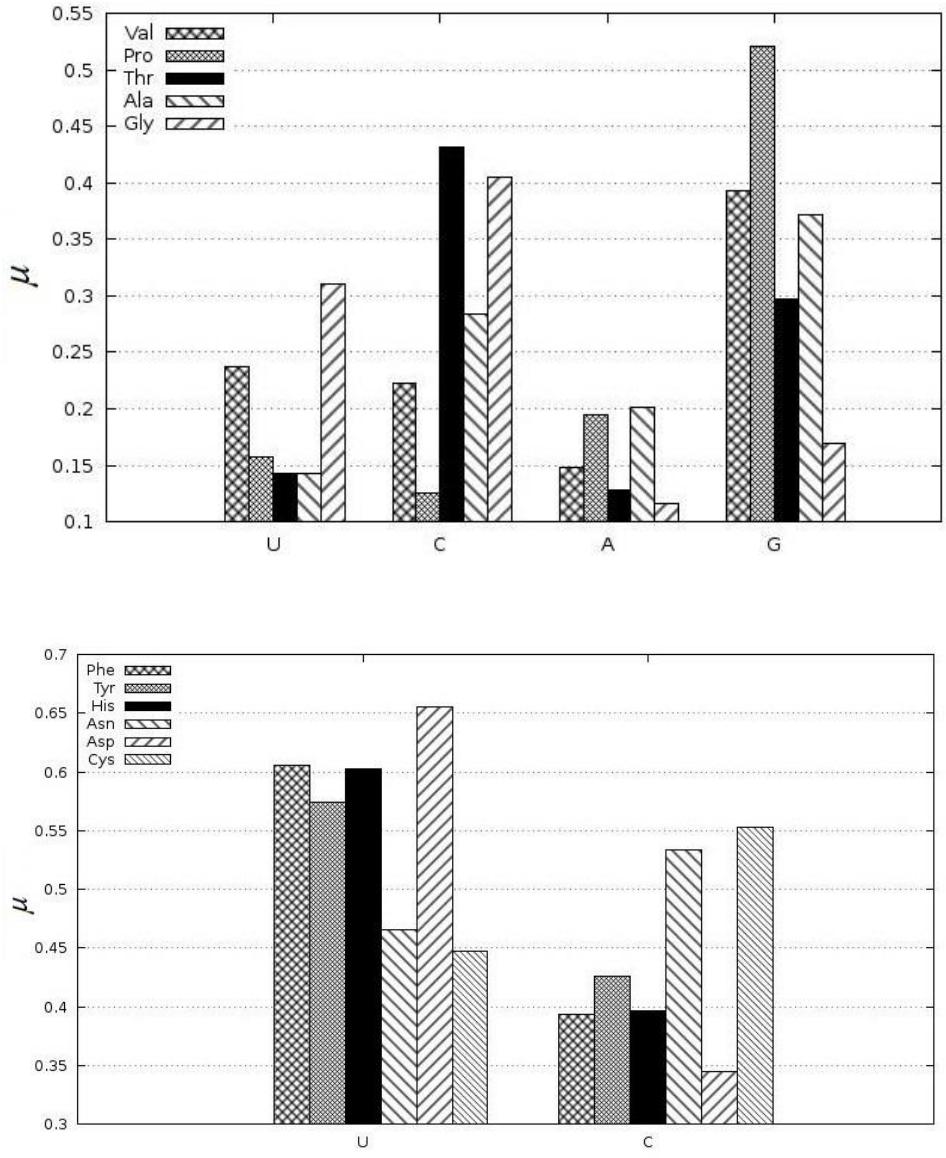

Figure 3. 4: Comparison of the μ parameters (y axis) in U+C two codon families, top, and the four codon families, bottom. In the x axis each letter shows the nucleotide in the third position of the codons in each amino acid.

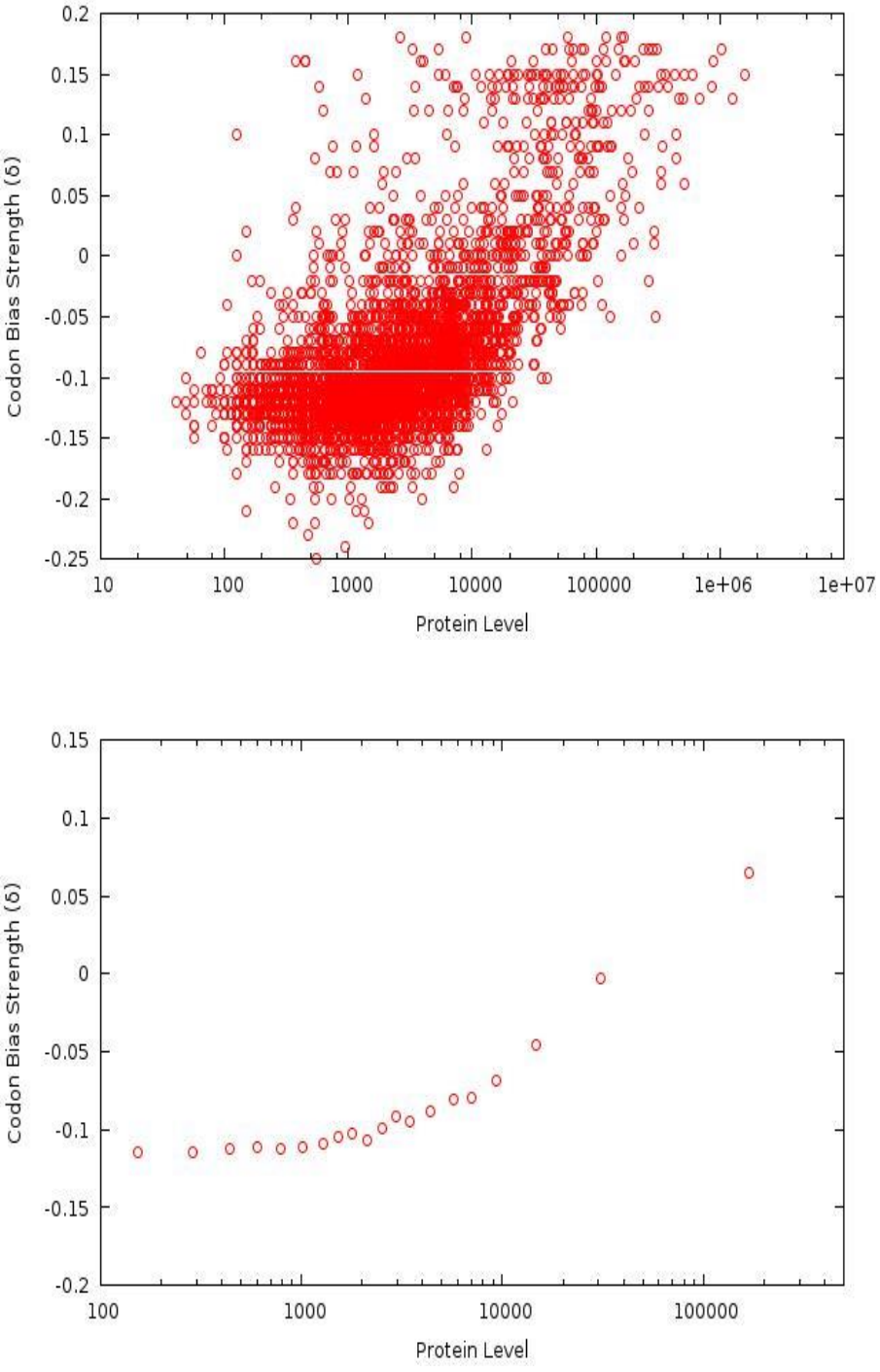

Figure 3. 5: This plot shows the relation between codon bias strength and protein level in yeast. Top - individual proteins; Bottom - binned into 40 bins.

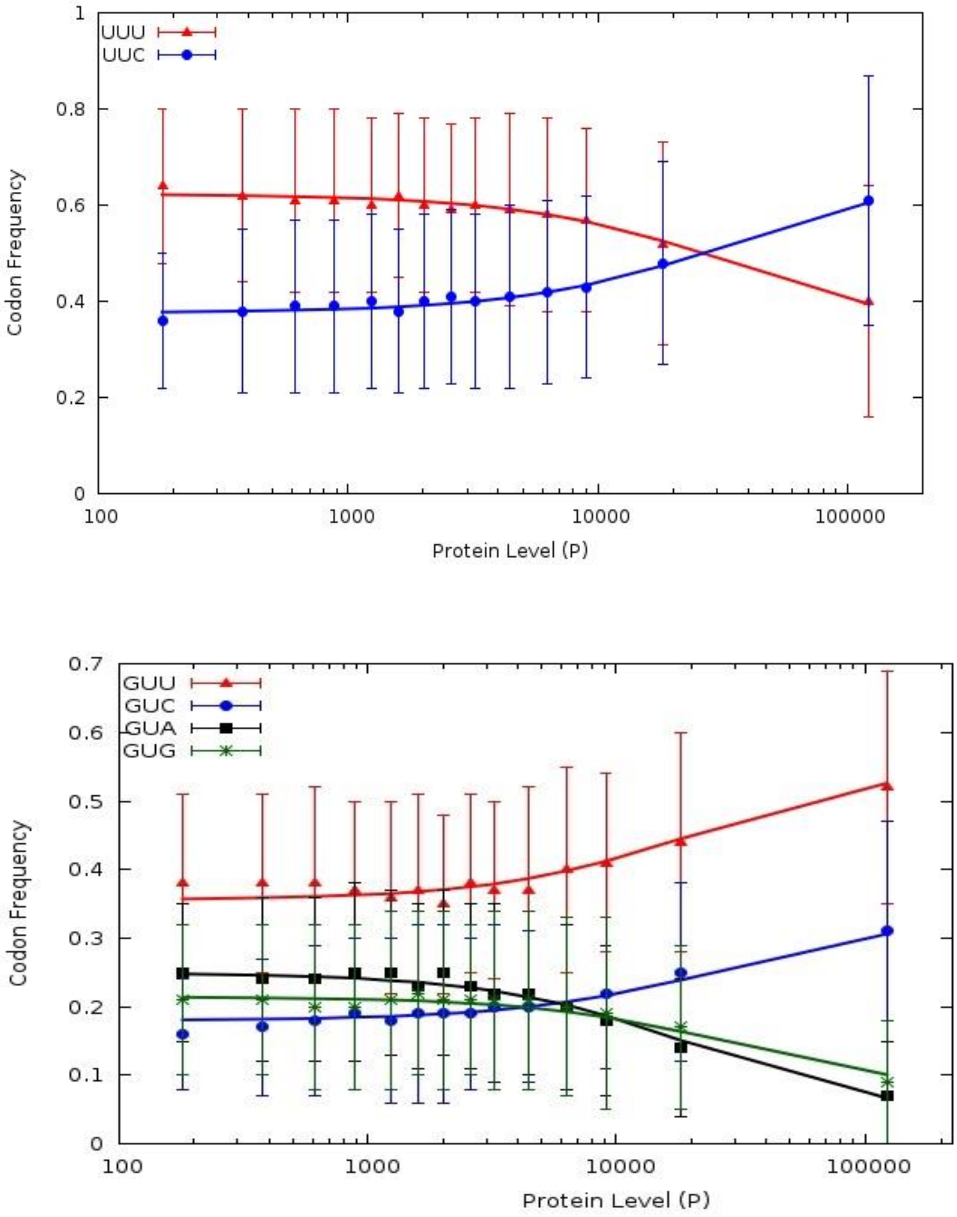

Figure 3.6: Codon frequency vs protein level for Phenylalanine, top, and Valine, bottom. The error bars shows one standard deviation in frequency in each bin.

Model	Selection function	Parameters	Ln L	ΔAIC
Standard	$S_{im} = k_i P_m$	41 μ + 41 k	-354592	0
Single point mutation	$S_{im} = k_i P_m$	3 μ + 41 k	-385849	+62438
Mutation dependent on 2nd position	$S_{im} = k_i P_m$	12 μ + 41 k	-368108	+26974
Restricted k (1)	$S_{im} = k_i P_m$	41 μ + 36 k	-354640	+86
Restricted k (2)	$S_{im} = k_i P_m$	41 μ + 39 k	-354608	+28
Logarithmic	$S_{im} = k_i \ln P_m$	41 μ + 41 k	-355031	+878
Power law	$S_{im} = k_i P_m^{\alpha}$	41 μ + 41 k +α	-353893	-1390
Saturating	$S_{im} = \frac{k_i P_i}{1 + P_i / P_{sat}}$	41 μ + 41 k + P_{sat}	-353746	-1692

Table 3.1: For this table we have binned 1018 genes in E.coli (for which the protein level is measured). Putting restrictions on mutation rates and selection coefficients would cause a significant loss in the information.

Model	Selection function	Ln L	ΔAIC
Standard	$S_{ig} = k_i P_g$	-224380	0.0
	$S_{ig} = k_i M_g$	-225139	+1518
	$S_{ig} = k_i P_g / M_g$	-225458	+2156
	$S_{ig} = k_i P_g / (M_g T_g)$	-225682	+2604
	$S_{ig} = k_i (P_g + a M_g)$	-224273	-214
Saturating	$S_{ig} = \frac{k_i P_g}{1 + P_g / P_{sat}}$	-223649	-1462
	$S_{ig} = \frac{k_i M_g}{1 + M_g / M_{sat}}$	-224694	+628
	$S_{ig} = k_i \left(\frac{P_g}{1 + P_g / P_{sat}} + a' \frac{M_g}{1 + M_g / M_{sat}} \right)$	-223602	-1556

Table 3.2: Comparison between different selection strength functions. The values for saturating functions which result in the highest likelihood are: $P_{sat} \approx 1.9 P_{average}$, $M_{sat} \approx 4.1 M_{average}$, $a \approx 0.5$ & $a' \approx 0.2$.

Model	Selection function	Parameters	Ln L	ΔAIC
Standard	$S_{im} = k_i P_m$	41 μ + 41 k	-2051334	0.0
Single point mutation	$S_{im} = k_i P_m$	3 μ + 41 k	-2103951	+105158
Mutation dependent on 2nd position	$S_{im} = k_i P_m$	12 μ + 41 k	-2092733	+82740
Saturating	$S_{im} = \frac{k_i P_i}{1 + P_i / P_{sat}}$	41 μ + 41 k	-2048949	-4770

Table 3.3: Comparison of different models for yeast.

Codon	Am.Acid	Φ^H	Φ^0	δ	tRNA Genes	μ	$100 \times k$
UUU	Phe	0.228	0.574	-0.922	0	0.606	-0.992
UUC	Phe	0.772	0.426	0.594	2	0.394	0
UUA	Leu	0.024	0.131	-1.688	1	0.137	-1.453
UUG	Leu	0.04	0.128	-1.158	1	0.139	-1.097
CUU	Leu	0.048	0.104	-0.764	0	0.105	-0.742
CUC	Leu	0.038	0.104	-1.004	1	0.105	-0.586
CUA	Leu	0.002	0.037	-2.901	1	0.04	-1.536
CUG	Leu	0.847	0.497	0.534	4	0.474	0
AUU	Ile	0.249	0.508	-0.713	0	0.538	-0.747
AUC	Ile	0.749	0.42	0.578	3	0.395	0
AUA	Ile	0.002	0.072	-3.446	0	0.067	-2.195
AUG	Met	1	1	0	8	1	0
GUU	Val	0.52	0.259	0.697	0	0.237	0
GUC	Val	0.078	0.216	-1.02	2	0.223	-0.755
GUA	Val	0.278	0.154	0.593	5	0.148	-0.19
GUG	Val	0.124	0.371	-1.098	0	0.393	-0.627
UCU	Ser	0.408	0.145	1.031	0	0.127	0
UCC	Ser	0.257	0.149	0.546	2	0.137	-0.179
UCA	Ser	0.031	0.123	-1.37	1	0.112	-1.263
UCG	Ser	0.013	0.154	-2.508	1	0.166	-1.389
CCU	Pro	0.142	0.158	-0.111	0	0.158	-0.566
CCC	Pro	0.011	0.124	-2.376	1	0.126	-1.516
CCA	Pro	0.126	0.191	-0.413	1	0.195	-0.579
CCG	Pro	0.72	0.527	0.313	1	0.521	0
ACU	Thr	0.468	0.166	1.037	0	0.143	0
ACC	Thr	0.442	0.435	0.016	2	0.431	-0.271
ACA	Thr	0.049	0.131	-0.981	1	0.128	-1.268
ACG	Thr	0.041	0.268	-1.87	2	0.297	-1.118
GCU	Ala	0.462	0.161	1.053	0	0.143	0
GCC	Ala	0.082	0.27	-1.196	2	0.284	-0.823
GCA	Ala	0.269	0.214	0.229	3	0.201	-0.353
GCG	Ala	0.188	0.356	-0.637	0	0.371	-0.533
UAU	Tyr	0.238	0.569	-0.873	0	0.574	-0.574
UAC	Tyr	0.762	0.431	0.571	3	0.426	0
CAU	His	0.299	0.572	-0.649	0	0.603	-0.921
CAC	His	0.701	0.428	0.493	1	0.397	0
CAA	Gln	0.197	0.347	-0.568	2	0.36	-0.669
CAG	Gln	0.803	0.653	0.207	2	0.64	0

AAU	Asn	0.123	0.451	-1.303	0	0.466	-0.96
AAC	Asn	0.877	0.549	0.469	4	0.534	0
AAA	Lys	0.715	0.766	-0.069	6	0.765	0
AAG	Lys	0.285	0.234	0.197	0	0.235	-0.201
GAU	Asp	0.359	0.627	-0.559	0	0.655	-0.664
GAC	Asp	0.641	0.373	0.543	3	0.345	0
GAA	Glu	0.763	0.69	0.101	4	0.684	0
GAG	Glu	0.237	0.31	-0.27	0	0.316	-0.302
UGU	Cys	0.316	0.445	-0.342	0	0.447	-0.418
UGC	Cys	0.684	0.555	0.209	1	0.553	0
UGG	Trp	1	1	0	1	1	0
CGU	Arg	0.683	0.38	0.586	4	0.353	0
CGC	Arg	0.302	0.4	-0.28	0	0.413	-0.584
CGA	Arg	0.003	0.064	-2.951	0	0.071	-2.304
CGG	Arg	0.005	0.098	-2.97	1	0.115	-2.524
AGU	Ser	0.047	0.151	-1.169	0	0.164	-1.646
AGC	Ser	0.245	0.277	-0.125	1	0.293	-0.896
AGA	Arg	0.007	0.037	-1.711	1	0.028	-2.267
AGG	Arg	0	0.021	-3.446	1	0.02	-3.244
GGU	Gly	0.621	0.338	0.61	0	0.31	0
GGC	Gly	0.359	0.404	-0.117	4	0.405	-0.338
GGA	Gly	0.006	0.108	-2.813	1	0.116	-1.678
GGG	Gly	0.013	0.151	-2.454	1	0.169	-1.366

Table 3.4: Table for the frequencies, δ values, number of tRNA genes, μ and the selection coefficients from the best fitted model for the codons, excluding the stop codons.

Chapter 4

Effect of mRNA secondary structure on codon usage in the beginning region of the gene sequences

4.1- Introduction:

Deviation in codon usage in the beginning region of the gene sequences has been reported in several studies (Bentele et al., 2013; Eyre-Walker & Bulmer, 1993; Gu et al., 2010; Tuller et al., 2010), but there's a lack of clear explanation for the phenomenon. In this study we try to investigate the relation between suppression of mRNA secondary structure and the reduction in codon adaptation in the first 10-15 codon positions of the gene sequences.

4.2- Materials and Method:

For this section we downloaded the *E. coli* genome sequence, 4141 protein genes, from the NCBI data base. Each gene was split into segments of 39 nucleotides (13 codons), using a sliding window where each window would be moved by two codons to produce the successive one. Mean free energy of the secondary structure in each segment was calculated using the program RNAfold [from ViennaRNA package (Lorenz et al., 2011)]

which reads in the sequence and finds the most stable structure that can be made and reports its mean free energy.

RNA secondary structure prediction through energy minimization is the most used function in the package. In this package three kinds of dynamic programming algorithms for structure prediction is provided: the minimum free energy algorithm of Zuker & Stiegler (1981), which yields a single optimal structure, the partition function algorithm of McCaskill (1990) which calculates base pair probabilities in the thermodynamic ensemble, and the suboptimal folding algorithm of Wuchty et al. (1999), which generates all suboptimal structures within a given energy range of the optimal energy.

The choice of 39 nucleotides is motivated by two facts, first that it is about the same number of codons in the beginning of the gene sequences which show a significant reduction in the δ value as compared to the rest of sequence, and folding in mRNA molecules is mainly short range and does not involve a long portion of the sequence, and also this is approximately the same size as the ribosome and if we want to argue about the effects of mRNA secondary structure in the translation initiation region of the mRNA molecules on the binding rate of the ribosomes to the mRNAs and translation initiation, it would reasonable to consider windows of such size. The data for gene expression level has been taken from Taniguchi et al., as described in chapter 3.

In order to look at the folding patter of the genes in *E. coli.* we split the gene sequences into segments of 39 nucleotides using a sliding window. Each window is made by moving the previous one by 6 nucleotides, the choice of fewer number of nucleotides does not change the pattern we observe. In this case the window covering the first 13

codons (starting from ATG) would be called the 1st codon window, and the window covering the range of 3rd-15th codon is called the 3rd codon window, and so on. This method, with different sliding intervals, has been used by others (Bentele et al., 2013; Tuller et al., 2010) as well.

4.3- A "Reduced Adaptation Region" in the beginning of the genes

As a measure of codon bias strength, we use δ, as described in chapter 1.When plotting the average δ for each codon position in *E. coli*, we can observe a significant decrease in δ value of the codons in the first 10-15 codons, Figure 4.1. This shows the existence of inefficiently translated codons in the beginning of the open reading frame. The same pattern was previously observed by (Tuller et al., 2010; Eyre-Walker & Bulmer, 1993). We will call this region the "reduced adaptation region (RAR)" for reference in future. To further explore the behavior of this phenomenon, we tried to see whether there is a relation between the severity of the RAR and the expression level of the genes. Since there's a positive correlation between the average δ value of each gene and its expression level, as explained in Chapter 3, we have plotted the average δ value in the beginning region of the genes vs the average codon bias strength of the whole gene sequence, Figure 4.2. It can be seen that the δ of the RAR increases roughly linearly with the δ for the rest of the gene, but δ of the RAR is lower than the δ for the rest of the gene at all δ values. This suggests that some factor that competes with selection for translational efficiency is acting in the RAR. The reduction in δ of the RAR relative to the rest of the gene becomes larger

as δ increases, which suggests that the competing factor is more in conflict with selection for translational efficiency in the genes with high δ and high expression level. This finding may imply the importance of the reduced adaptation region for protein production.

4.4- mRNA secondary structure and RAR

Folding pattern in E. coli genes

As it can be seen in Figure 4.3, there's a clear reduction in the folding free energy of the mRNA sequences up to the 5^{th} codon window. Tuller et al., have shown that the reduction in folding in the 1^{st} codon window compared to the folding energies in the rest of the gene sequence is selected for (Tuller et al., 2010). One interesting feature that can be detected in this plot is the significant increase in the strength of folding (i.e more negative values of folding free energy) in the 7^{th}-11^{th} codon windows, which covers structure in the interval of 7^{th}-24^{th} codon,. This strong folding might exist in order to prevent formation of strong secondary structures in the translation initiation region of the genes (Tuller et al., 2010).It is interesting to notice that the windows where the folding energies are significantly different from the middle region (i.e. windows beyond codon 15) covers the same codons as the region showing low translation efficiency, when plotting codon bias strength of codons versus their position in the gene sequence.

Generation of synonymous random genes and selection for weak folding

In order to see whether the secondary structure of the mRNA molecules in the beginning of ORF plays a role in shaping the codon bias in the RAR, we generated synonymous random sequences having the same resultant amino acid chains and codon

frequencies equal to φ^0, frequency of codons when averaged over the whole genome. And also randomized sequences with frequencies equal to that of highly expressed genes, φ^H. For each gene, a set of 1000 such sequences were produced. In each set of randomized sequences the folding energy profile, as discussed in the method section, was measured and in the set with frequencies equal to φ^0, the ones with the weakest secondary structure, least 1% and 0.1%, were chosen. Figure 4.4 shows these values for the real and randomized sequences and the ones with least 0.1% folding energy in the first codon window. For the selected sequences, which show weak secondary structure in the beginning region, the translation efficiency profile of the codons, codon bias strength of codons vs their position in the sequence, was plotted and compared with the translation efficiency profile observed in the real genes,

As it is evident from Figure 4.4, selecting for weak folding in the beginning region of the randomly generated genes also leads to a reduction in δ in the same region, as is observed for real genes. This might strengthen the idea that appearance of the RAR in the translation efficiency profile of the codons in the first 10-15 codons is a side effect of suppressing the secondary structure in this region.

But looking at the folding energy pattern of the randomized sequences with φ^0 frequencies, we see that for these sequences before any selection for weak folding in the first codon window, even though there is no bias in codon usage in different positions of the genes, the blue curve in Figure 4.4, the folding free energy in the first 13 codons is more or less the same as what we observe in real genes, the blue curve in Figure 4.3. When we select for the sequences with weak secondary structures in the first codon window, the

chosen sequences show folding free energies which are much less negative than what we observe in real genes in the beginning region, -2.0 kcal/mol compared to -5.2 kcal/mol. We have also generated synonymous sequences having codon frequencies equal to that of highly expressed genes, φ^H, and measured the folding energy pattern of these sequences as well. As it can be seen, the brown curve in Figure 4.3, even substituting the inefficient codons in real sequences with efficient ones, while keeping the coding sequence the same, would not result in a much different folding free energy of the first codon window although it would change the δ profile a lot. In other words it is possible to use the efficient codons, in terms of translation efficiency, in all of the positions in the gene sequence without appreciably changing the strength of secondary structure in the beginning of the ORF. The secondary structure seems to be mostly governed by the first two codon positions, the positions which (except for the six codon families) don't change when codons are synonymously substituted, not the third position.

Bentele et al., (Bentele et al., 2013), have reported the reduction in the GC content, in all three nucleotide positions of the codons, at the beginning of the genes in *E. coli*. This observation along with our results would imply that the selection for weak folding in the beginning region of the mRNA sequences is strong enough to not only select for synonymous codons that have low GC levels but would also result in sequences that code for amino acids with low GC content averaged over their synonymous codons in the first 10-15 codon positions. In order to test this argument, and since the effect of reduction in adaptation of codons is more significant in highly expressed genes, we focused on the highly expressed genes and calculated the amount by which each amino acid goes up or

down in the positions 2^{nd}-13^{th}, since the start codon is almost always the same. We then calculated the GC content of each amino acid averaged over its codons. We both calculated the GC content in the first two nucleotide positions and all three nucleotide positions of the codons separately to see which one is more correlated with the change in frequency of each amino acid. As it can be clearly seen from the Figure 4.5, a negative correlation exists between the usage of an amino acid in the beginning region of the genes and the GC content in the first two nucleotide positions or the total GC content of its codons, R^2=0.30 & R^2=0.34 respectively. This means that higher the GC content of the synonymous codons coding for a specific amino acid is, the more its usage is suppressed in the beginning of the sequence. Then we tried to see whether using the amino acids that increase in frequency in the beginning region of the sequence, would automatically result in a reduction in adaptation. We define the average adaptation of an amino acid:

$$\delta_j = \sum_{i \in aa} \varphi_{ij} \times \delta_i \qquad (21)$$

where the indices *j* and *i* indicate the amino acid *j* and the codon *i* coding for that amino acid respectively, and the summation is over the synonymous codons which code for the amino acid *j*. By putting the frequencies equal to that of highly expressed genes we would get the average adaptation of each amino acid in the highly expressed genes. These values are plotted against the amount by which each amino acid increases or decreases in frequency in the reduced adaptation region compared to the rest of the gene sequence, Figure 4.6, and no strong correlation could be observed, R^2=0.003. In other words, the

amino acids which increase in frequency in the reduced adaptation region, would not automatically result in a reduction in the translation efficiency profile of the genome.

But one difference between the generated sequences and the real ones is the reduction of the strength of folding in the 7^{th}-11^{th} codon windows in random sequences when compared to the real sequences. In the real sequences the average folding free energy in these windows is -8.07 kcal/mol when the average folding in the windows after the 15^{th} codon one is -7.43 kcal/mol, whereas in the random sequences generated with φ^0frequencies, the values are -7.51 and -6.97 kcal/mol respectively, and for the ones generated with φ^H frequencies the values are -7.32 and -6.95 kcal/mol respectively. This fact would motivate further investigation in the relation between folding energy and the codon bias in the beginning region of the gene sequences and also translation initiation.

4.5- mRNA secondary structure and the gene expression level:

In the analysis of the data of Taniguchi et al. in Chapter 3 we showed that the ratio *P/(MT)* for a gene is equal to the rate of protein production per mRNA. We also showed that the rate of protein production per mRNA is equal to the binding rate of ribosomes at the beginning of the gene. In this chapter we have supposed that reduction in secondary structure is important to allow efficient ribosome binding, and this is what is responsible for the observed changes in folding free energy and codon usage at the beginnings of genes. To further test this idea, we investigated the relationship between *P/MT* in the data of Taniguchi et al and the folding of the mRNAs for these genes. For this task we binned

the genes for which the protein production rate is measured by Taniguchi et al. according to the folding energy in the translation initiation region. The result show a weak correlation (R^2=0.007) between the folding energy in the first codon window and the protein production rate over each mRNA, Figure 4.7. Number of protein molecules produced per mRNA and the total protein level show even less correlation, R^2 = 0.006 & R^2 = 0.005 respectively, with the folding free energy of the first codon window.

As it can be seen from the plots, no strong correlation between the folding energy in the beginning region of mRNA and different expression measures can be seen. Over a wide range of change in folding energy, all gene expression measures almost stay the same. The weak correlation between folding in the first 13 codons and expression level, P/M, in *E. coli* has also been reported by Tuller et al., (Tuller et al., 2010).

One last thing to investigate is whether the difference in average folding energies of the 7^{th}-11^{th} codon windows, and the folding energy in the first codon window has any effect on or correlation with the production level. The motivation would be that in the genes where the sequence is capable of producing strong secondary structure in the mentioned region, the chance of strong folding in the ribosomal binding cite would decrease. In other words the actual value of folding free energy in the initiation region may not be important, as long as there is a segment ahead with higher chance of forming a strong secondary structure. For this matter we selected the genes which are long enough to at least include one window in the interval of 7^{th}-11^{th} codon windows, and calculated the difference in folding energy of the first window and the average folding energy of the 7^{th}-11^{th} codon windows. Protein production rates of 575 genes, for which the information is

provided by Taniguchi et al. (Taniguchi et al., Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells, 2010), was plotted against this value, Figure 4.8. As it can be seen there is still no strong correlation ($R^2 = 0.003$) between the expression level of each individual gene and the difference in the folding in the first window and the average folding energy of the 7^{th}-11^{th} codon windows. But if we look at the folding energy values in the genes, we see that only 23% of the genes have stronger folding in the first codon window compared to the average folding in the 7^{th}-11^{th} codon windows. The average of this difference, the difference between the folding free energy in the first codon window and the average folding free energy in the 7^{th}-11^{th} windows, over all the genes is 2.60 kcal/mol, with standard deviation equal to 3.82 kcal/mol.

4.6- Conclusion and Discussion:

Existence of a secondary structure in mRNA could interfere with protein production in many ways, but the translation initiation could still occur if the detection of the start codon is not altered, Figure 4.9. Here we have focus on the secondary structures within the ORF. By providing synonymous random sequences, sequences with the same amino acid chain but frequencies equal to φ^0, and selecting for the ones with weak folding in the beginning region we observe that a translation efficiency profile much similar to that of real genes would appear, but the selected sequences show folding free energies much different from that of real sequences, Figure 4.3 & *Figure 4.4*. The other feature observed

is that in the randomized sequences even thou the reduced adaptation region disappears, there is still a clear reduction in the folding free energy in the first 13 codons compared to the rest of the gene. To see whether the selection for weak secondary structures in the beginning of the gene sequences is strong enough to select for sequences that code for amino acids with overall low GC content, we calculated the amount of change in the frequency of each amino acid in the first 13 codon position of the gene sequences and the rest of the gene in highly expressed genes, since they are assumed to be under strong selection for proper translation. These values were plotted against the GC content of the synonymous codons coding for each amino acid and a clear tendency for choosing amino acids with overall low GC content among their synonymous codons in the first 13 codon positions of the genes was observed. One explanation for the existence of the reduced adaptation region would be that the amino acids that are increased in frequency in the beginning of the gene sequences, are the ones that, on average, show a low codon adaptation. We plotted the amount of change of each amino acid against its average codon adaptation, equation 1, but no strong correlation could be observed, Figure 4.6. We further observe that the protein production rate over each mRNA sequence in *E. coli* does not show any strong correlation between the folding free energy in the beginning of the ORF, Figure 4.7 & *Figure 4.8*.

In conclusion, our results clearly indicate that there is a selection for suppression of mRNA secondary structure in the beginning of the gene sequences and also there's a significant reduction in the translation efficiency in the codons used in the first 10-15 codon positions of the sequences, and as the expression level of the genes increase, even

thou the whole sequence is under selection for using frequent codons, the difference between the average adaptation of the codons in the beginning and the rest of the gene increases. But our method for calculating the strength of secondary structure in this region does not show any significant relation between the suppression of mRNA secondary structure in the first 13 codons and the translation efficiency profile observed in the genes. This observation could be due to the fact that we have not considered the Shine–Dalgarno (SD) sequence upstream the start codon and have only focused on the ORF. For further investigation on this matter one could look at broader window lengths including both the ORF and SD region, and see whether the translation efficiency profile could be resulted by selection of the sequences which reduce the strength of secondary structures in this region, or the ones that form structures which does not include the start codon.

4.7- Figures and Tables

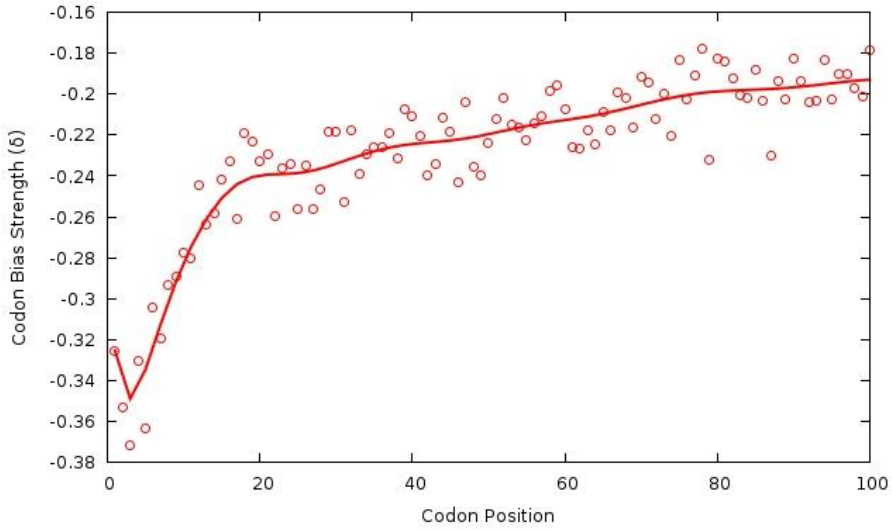

Figure 4.1: Translation efficiency profile, δ vs codon position, of E.coli protein genes. There's a region of reduced adaptiveness in the beginning of the genes, first 10-15 codons.

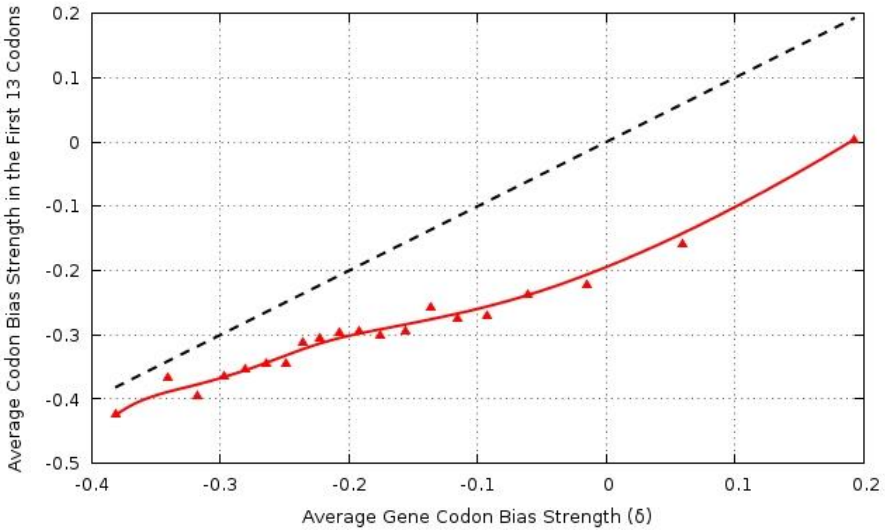

Figure 4.2: Plot of average δ value of the beginning region, first 13 codons, vs the average δ of the whole gene for 4141 E. coli protein genes.

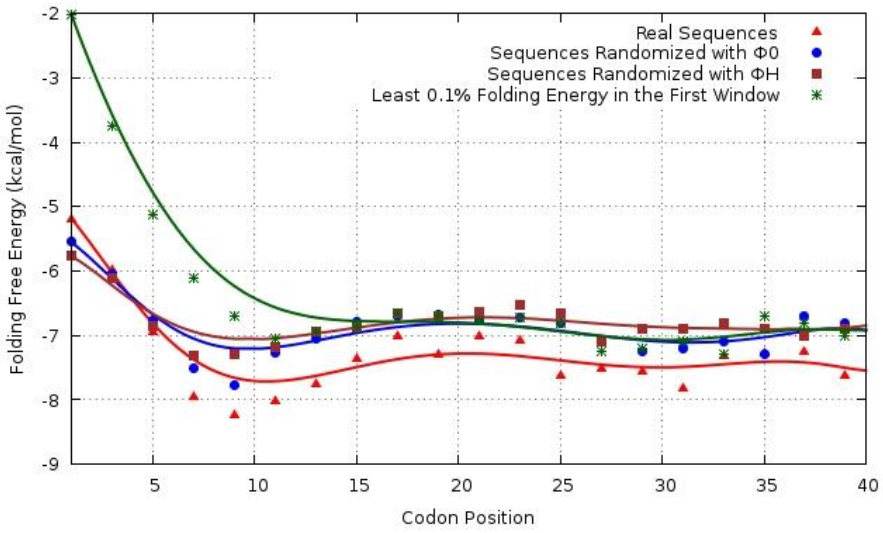

Figure 4.3: Folding free energy profile for the protein genes of E. coli. Codon position indicates the position of the codon by which the window starts, starting from 1, for the start codon, for the first window.

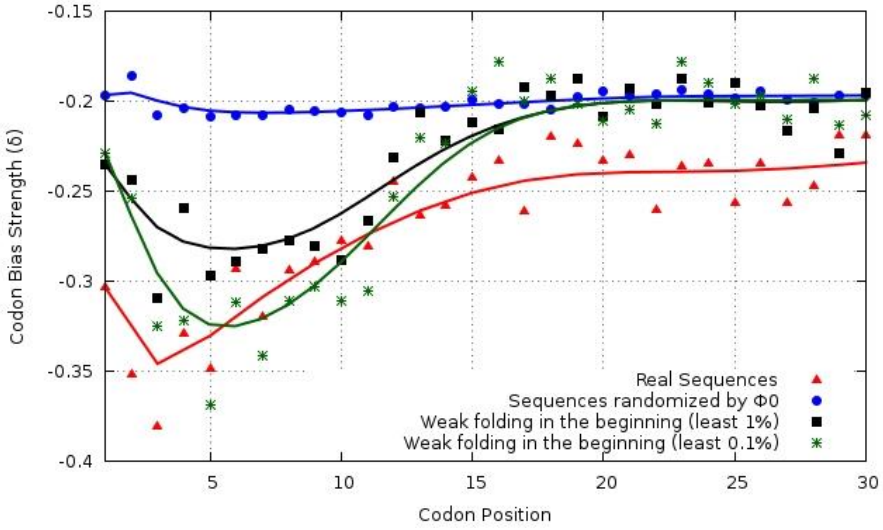

Figure 4.4: δ vs codon position for real genes and the randomized sequences. Sequences randomized by φ^0 frequencies, blue solid line, show no bias in codon usage in the beginning region compared to the rest of the sequence, and as we increase the selection for weak folding in the first 13 codons among the randomized sequences, black and green solid lines, we observe the appearance of a region of reduced adaptation.

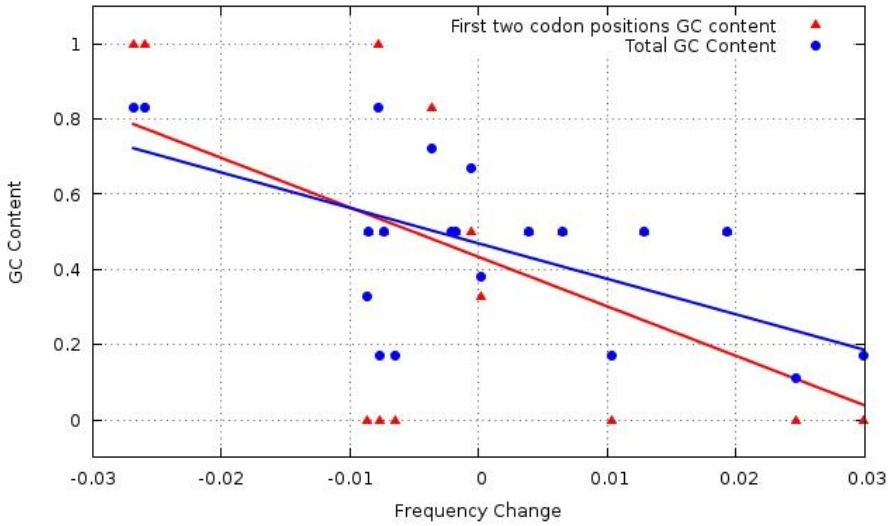

Figure 4.5: GC content of the amino acids vs the amount by which they increase or decrease in the beginning region compared to the rest of the gene, in highly expressed genes. Blue markers show the GC content of each amino acid average over all three positions of its codons, and the red markers indicate the GC content averaged over only the first two nucleotide positions of the codons coding for one specific amino acid.

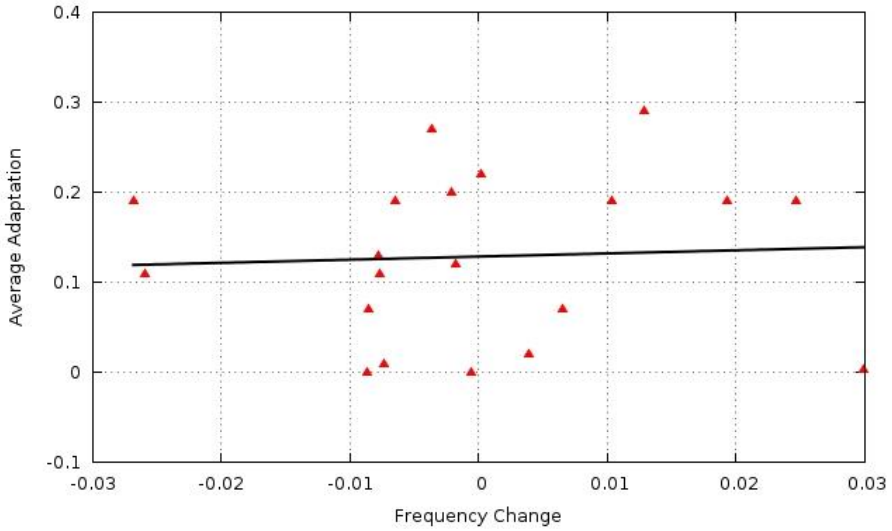

Figure 4.6: The average adaptation of each amino acid vs the amount by which it increases or decreases in the first 13 codons, in highly expressed genes.

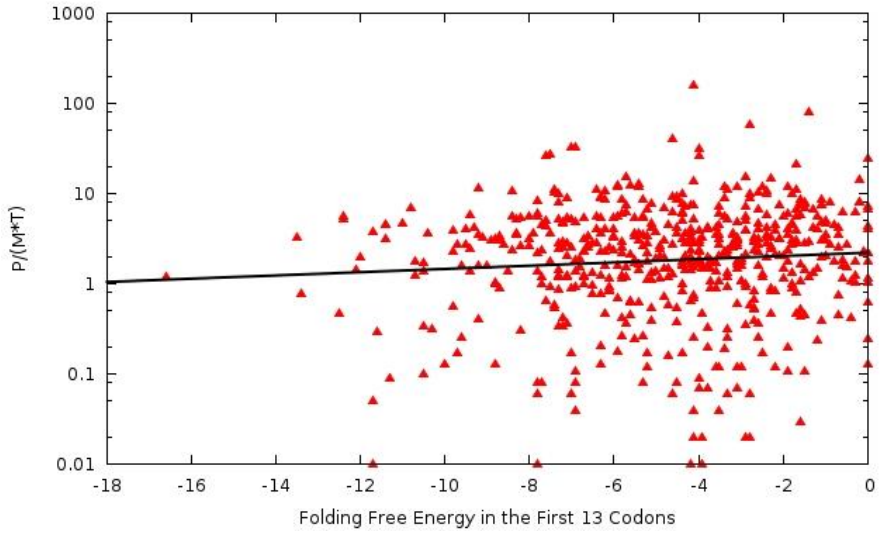

Figure 4.7: Protein production rate per mRNA molecule vs folding free energy in the beginning region of the genes. There is a very weak correlation, R^2=0.007, observed between these two parameters.

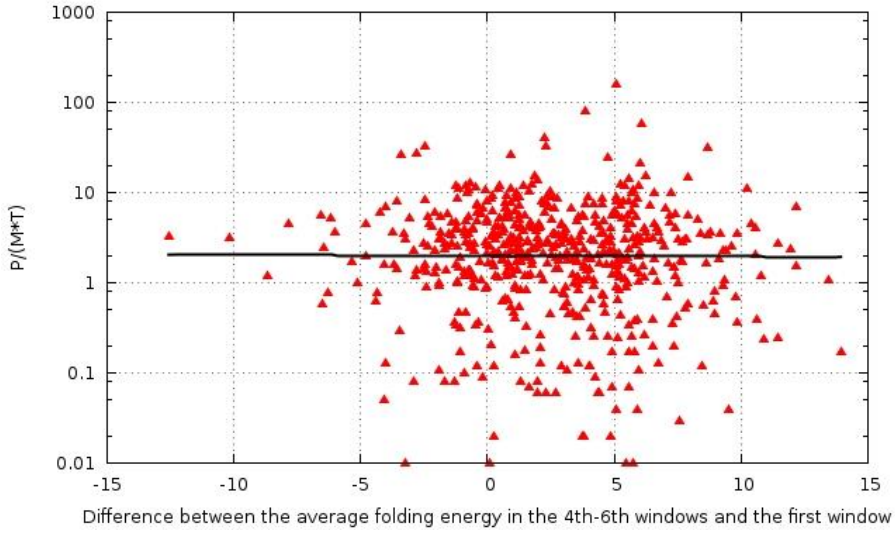

Figure 4.8: Protein production rate per mRNA molecule vs the difference between average folding free energy in the 7^{th}-11^{th} codon windows and the first 13 codons. No significant correlation can be observed between the two parameters.

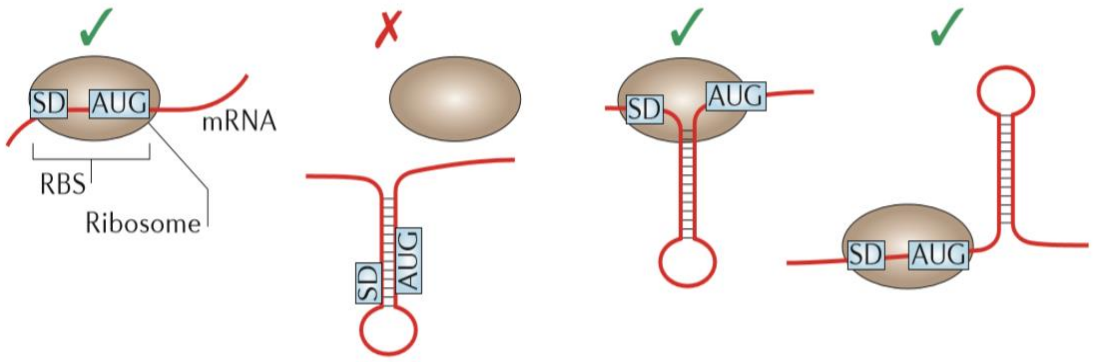

Figure 4.9: Formation of mRNA secondary structure in the ribosome binding site (RBS) could usually inhibit translation initiation. However, initiation can occur when the structured element is positioned between the Shine–Dalgarno sequence (SD) and the start codon (AUG) *(Nivinskas et al., 1999)*. [Photo taken from *(Plotkin & Kudla, 2010)*]

Chapter 5

References

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In S. N. York, *Selected Papers of Hirotugu Akaike* (pp. 199-213).

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

Alff-Steinberger, C. (1969). The genetic code and error transmission. *Proceedings of the National Academy of Sciences, 64*(2), 584-591.

Amirnovin, R. (1997). An analysis of the metabolic theory of the origin of the genetic code. *Journal of molecular evolution, 44*(5), 473-476.

Andersson, S. G., & Sharp, P. M. (1996*a*). Codon usage in the Mycobacterium tuberculosis complex. *Microbiology, 142*(4), 915-925.

Andersson, S. G., & Sharp, P. M. (1996*b*). Codon usage and base composition inRickettsia prowazekii. *Journal of molecular evolution, 42*(5), 525-536.

Bennetzen, J. L., & Hall, B. D. (1982). Codon selection in yeast. *Journal of Biological Chemistry, 257*(6), 3026-3031.

Bentele, K., Saffert, P., Rauscher, R., Ignatova, Z., & Blüthgen, N. (2013). Efficient translation initiation dictates codon usage at gene start. *Molecular systems biology, 9*(1).

Bergmann, J. E., & Lodish, H. F. (1979). A kinetic model of protein synthesis. Application to hemoglobin synthesis and translational control. *Journal of Biological Chemistry, 254*(23), 11927-11937.

Bettany, A. J., Moore, P. A., Cafferkey, R., Bell, L. D., Goodey, A. R., Carter, B. L., & Brown, A. J. (1989). 5′-Secondary structure formation, in constrast to a short string of non-preferred codons, inhibits the translation of the pyruvate kinase mRNA in yeast. *Yeast, 5*(3), 187-198.

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

Bulmer, M. (1988). Codon usage and intragenic position. *Journal of theoretical biology, 133*(1), 67-71.

Burns, D. M., & Beacham, I. R. (1985). Rare codons in E. coli and S. typhimurium signal sequences. *FEBS letters, 189*(2), 318-324.

Chamary, J. V., Parmley, J. L., & Hurst, L. D. (2006). Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nature Rev. Gene, 7*(2), 98–108.

Chen, S. L., Lee, W., Hottes, A. K., Shapiro, L., & McAdams, H. H. (2004). Codon usage between genomes is constrained by genome-wide mutational processes. *Natl. Acad. Sci, 101*(10), 3480-3485.

Chou, T., & Lakatos, G. (2004). Clustered bottlenecks in mRNA translation and protein synthesis. *Physical review letters, 93*(19).

Crick, F. H. (1988). What mad pursuit: A personal view of scientific discovery. New York: Basic books.

Crick, F. H. (1968). The origin of the genetic code. *Journal of molecular biology, 38*(3), 367-379.

Curran, J. F., & Yarus, M. (1989). Rates of aminoacyl-tRNA selection at 29 sense codons in vivo. *Journal of molecular biology, 209*(1), 65-77.

Drummond & Wilke. (2008). Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell, 134*(2), 341-352.

de Smit, M. H., & Van Duin, J. (1990). Secondary structure of the ribosome binding site determines translational efficiency: a quantitative analysis. *Proceedings of the National Academy of Sciences, 87*(19), 7668-7672.

Di Giulio, M. (1997). On the origin of the genetic code. *Journal of theoretical biology, 187*(4), 573-581.

Dong, J., Schmittmann, B., & Zia, R. K. (2007). Towards a model for protein production rates. *Journal of Statistical Physics, 128*(1-2), 21-34.

dos Reis, M., Savva, R., & Wernisch, L. (2004). Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic acids research, 32*(17), 5036-5044.

Duret, L., & Mouchiroud, D. (1999). Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. *Proceedings of the National Academy of Sciences, 96*(8), 4482-4487.

Duret, L. (2002). Evolution of synonymous codon usage in metazoans. *Current opinion in genetics & development, 12*(6), 640-649.

Epstein, C. J. (1966). Role of the amino-acidcode'and of selection for conformation in the evolution of proteins. *Nature, 210*, 25-28.

Eskesen, S. T., Eskesen, F. N., & Ruvinsky, A. (2004). Natural selection affects frequencies of AG and GT dinucleotides at the 5′ and 3′ ends of exons. *Genetics, 167*(1), 543–550.

Eyre-Walker, A., & Bulmer, M. (1993). Reduced synonymous substitution rate at the start of enterobacterial genes. *Nucleic acids research, 21*(19), 4599-4603.

Eyre-Walker, A., & Bulmer, M. (1995). Synonymous substitution rates in enterobacteria. *Genetics, 140*(4), 1407-1412.

Freeland, S. J., & Hurst, L. D. (1998). The genetic code is one in a million. *Journal of molecular evolution, 47*(3), 238-248.

Ghaemmaghami, S., Huh, W. K., Bower, K., Howson, R. W., Belle, A., Dephoure, N., ... & Weissman, J. S. (2003). Global analysis of protein expression in yeast. *Nature, 425*(6959), 737-741.

Gouy, M., & Gautier, C. (1982). Codon usage in bacteria: correlation with gene expressivity. *Nucleic acids research, 10*(22), 7055-7074.

Grantham, R., Gautier, C., Gouy, M., Mercier, R., & Pave, A. (1980). Codon catalog usage and the genome hypothesis. *Nucleic Acids Research, 8*(1), 197.

Grantham, R., Gautier, C., Gouy, M., Jacobzone, M., & Mercier, R. (1981). Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Research, 9*(1), 213.

Grantham, R. (1980). Working of the genetic code. *Trends in Biochemical Sciences , 5*(12), 327-331.

Greulich, P., & Schadschneider, A. (2008). Phase diagram and edge effects in the ASEP with bottlenecks. *378*(8), 1972-1986.

Gribskov, M., Devereux, J., & Burgess, R. R. (1984). The codon preference plot: graphic analysis of protein coding sequences and prediction of gene expression. *Nucleic acids research, 12*(1), 539-549.

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 computational biology, 6*(2), e1000664.

Haig, D., & Hurst, L. D. (1991). A quantitative measure of error minimization in the genetic code. *Journal of molecular evolution, 33*(5), 412-417.

Hershberg, R., & Petrov, D. A. (2008). Selection on Codon Bias. *Annual Review of Genetics, 42*, 287-299.

Higgs, P. G., & Ran, W. (2008). Coevolution of Codon Usage and tRNA Genes Leads to Alternative Stable States of Biased Codon Usage. *Mol Biol Evol, 25*(11), 2279-2291.

Higgs, P. G. (2009). A four-column theory for the origin of the genetic code: tracing the evolutionary pathways that gave rise to an optimized code. *Biology direct, 4*(1), 16.

Hinegardner, R. T., & Engelberg, J. (1963). Rationale for a universal genetic code. *Science, 142*(3595), 1083-1085.

Holley, R. W., Apgar, J., Everett, G. A., Madison, J. T., Marquisee, M., Merrill, S. H., ... & Zamir, A. (1965). Structure of a ribonucleic acid. *Science, 147*(3664), 1462-1465.

Ikemura, T. (1981). 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. *Journal of molecular biology, 151*(3), 389-409.

Ikemura, T. (1985). Codon usage and tRNA content in unicellular and multicellular organisms. *Molecular biology and evolution, 2*(1), 13-34.

Ingolia, N. T., Ghaemmaghami, S., Newman, J. R., & Weissman, J. S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science, 324*(5924), 218-223.

Kanaya, S., Kinouchi, M., Abe, T., Kudo, Y., Yamada, Y., Nishi, T., ... & Ikemura, T. (2001). Analysis of codon usage diversity of bacterial genes with a self-organizing map (SOM): characterization of horizontally transferred genes with emphasis on the E. coli O157 genome. *Gene, 276*(1), 89-99.

Kanaya, S., Yamada, Y., Kinouchi, M., Kudo, Y., & Ikemura, T. (2001). Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. *J. Mol. Biol, 53*(4-5), 290-298.

Khorana, H. G., Büuchi, H., Ghosh, H., Gupta, N., Jacob, T. M., Kössel, H., ... & Wells, R. D. (1966). Polynucleotide synthesis and the genetic code. *Cold Spring Harbor Symposia on Quantitative Biology.31*, pp. 39-49. Cold Spring Harbor Laboratory Press.

Knight, R. D., Freeland, S. J., & Landweber, L. F. (2001) a. Rewiring the keyboard: evolvability of the genetic code. *Nature Reviews Genetics, 2*(1), 49-58.

Knight, R. D., Freeland, S. J., & Landweber, L. F. (2001) b. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biology, 2*(4), research0010.

Kolmsee & Hengge. (2011). Rare codons play a positive role in the expression of the stationary phase sigma factor RpoS (σS) in Escherichia coli. *RNA biology, 8*(5), 913-921.

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.

Lengyel, P., Speyer, J. F., & Ochoa, S. (1961). Synthetic polynucleotides and the amino acid code. *Proceedings of the National Academy of Sciences of the United States of America, 47*(12).

Lengyel, P., Speyer, J. F., Basilio, C., & Ochoa, S. (1962). Synthetic polynucleotides and the amino acid code, III. *48*(2).

Lorenz, R., Bernhart, S. H., Zu Siederdissen, C. H., Tafer, H., Flamm, C., Stadler, P. F., & Hofacker, I. L. (2011). ViennaRNA Package 2.0. *Algorithms for Molecular Biology, 6*(1).

Mathews, M., Sonenberg, N., & Hershey, J. W. (Eds.). (2007). *Translational control in biology and medicine* (Vol. 48). New York: CSHL Press.

McCaskill, J. (1990). The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers, 29*(6-7), 1105-1119.

McLachlan, A. D., Staden, R., & Boswell, D. R. (1984). A method for measuring the non-random bias of a codon usage table. *Nucleic acids research, 12*(24), 9567-9575.

Mitarai et al. (2008). Ribosome collisions and translation efficiency: optimization by codon usage and mRNA destabilization. *Journal of molecular biology, 382*(1), 236-245.

Nirenberg, M. W., & Matthaei, J. H. (1961). The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides. *Proceedings of the National Academy of Sciences, 47*(10), 1588-1602.

Nirenberg, M. W., Jones, O. W., Leder, P., Clark, B. F. C., Sly, W. S., & Pestka, S. (1963). On the coding of genetic information. *Cold Spring Harbor Symposia on Quantitative Biology.29*, pp. 549-557. Cold Spring Harbor Laboratory Press.

Plotkin, J. B., & Kudla, G. (2010). Synonymous but not the same: the causes and consequences of codon bias. *Nature, 12*(1), 32-42. doi:10.1038/nrg2899

Qin, H., Wu, W. B., Comeron, J. M., Kreitman, M., & Li, W. H. (2004). Intragenic spatial patterns of codon usage bias in prokaryotic and eukaryotic genomes. *Genetics, 168*(4), 2245-2260.

Ran, W., & Higgs, P. G. (2010). The influence of anticodon–codon interactions and modified bases on codon usage bias in bacteria." Molecular biology and evolution. *27*(9), 2129-2140.

Ran, W., & Higgs, P. G. (2012). Contributions of Speed and Accuracy to Translational Selection in Bacteria. *PLoS ONE, 7*(12), e51652.

Sengupta, S., Yang, X., & Higgs, P. G. (2007). The mechanisms of codon reassignments in mitochondrial genetic codes. *Journal of molecular evolution, 64*(6), 662-688.

Shah, P., & Gilchrist, M. A. (2011). Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and genetic drift. *Proceedings of the National Academy of Sciences, 108*(25), 10231-10236.

Shah, P., Ding, Y., Niemczyk, M., Kudla, G., & Plotkin, J. B. (2013). Rate-Limiting Steps in Yeast Protein Translation. *Cell, 153*(7), 1589-1601.

Shapiro, B. A., & Zhang, K. (1990). Comparing multiple RNA secondary structures using tree comparisons. *Computer applications in the biosciences: CABIOS, 6*(4), 309-318.

Sharp, P. M., & Li, W. H. (1987). The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. *Molecular Biology and Evolution, 4*(3), 222-230.

Sharp, P. M., Bailes, E., Grocock, R. J., Peden, J. F., & Sockett, R. E. (2005). Variation in the strength of selection codon usage bias among bacteria. *Nucleic acids research, 33*(4), 1141-1153.

Sharp, P. M. (1986). Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic acids research, 14*(13), 5125-5143.

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 research, 15*(3), 1281-1295.

Shaw, L. B., Zia, R. K. P., & Lee, K. H. (2003). Totally asymmetric exclusion process with extended objects: a model for protein synthesis. *Physical Review E, 68*(2).

Shaw, L. B., Kolomeisky, A. B., & Lee, K. H. (2004). Local inhomogeneity in asymmetric simple exclusion processes with extended objects. *Journal of Physics A: Mathematical and General, 37*(6).

Sonneborn, T. M. (1965). Degeneracy of the genetic code: extent, nature, and genetic implications. *Evolving genes and proteins*, 377-397.

Sørensen, M. A., Kurland, C. G., & Pedersen, S. (1989). Codon usage determines translation rate in Escherichia coli. *Journal of molecular biology, 207*(2), 365-377.

Speyer, J. F., Lengyel, P., Basilio, C., & Ochoa, S. (1962). Synthetic polynucleotides and the amino acid code, II. *Proceedings of the National Academy of Sciences of the United States of America, 48*(1).

Stoletzki, N., & Eyre-Walker, A. (2007). Synonymous codon usage inEscherichia coli: selection for translational accuracy. *Mol. Biol. Evol, 24*(2), 374-381.

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.

Taylor, F. J. R., & Coates, D. (1989). The code within the codons. *Biosystems, 22*(3), 177-187.

Thanaraj, T. A., & Argos, P. (1996). Ribosome-mediated translational pause and protein domain organization. *Protein Science, 5*, 1594–1612.

Trotta E. (2013). Selection on codon bias in yeast: a transcriptional hypothesis. *Nucleic Acids Research*, gkt740.

Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., ... & Pilpel, Y. (2010). An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell, 141*(2), 344-354.

Tuller, T., Waldman, Y. Y., Kupiec, M., & Ruppin, E. (2010). Translation efficiency is determined by both codon bias and folding energy. *Proceedings of the National Academy of Sciences, 107*(9), 3645-3650.

Warnecke, T., & Hurst, L. D. (2007). Evidence for a trade-off between translational efficiency and splicing regulation in determining synonymous codon usage in Drosophila melanogaster. *Molecular biology and evolution, 24*(12), 2755–2762.

Watson, J. D., & Crick, F. H. C. (1953). A structure for deoxyribose nucleic acid. *Nature, 421*(6921), 397-3988.

Watts, J. M., Dang, K. K., Gorelick, R. J., Leonard, C. W., Bess Jr, J. W., Swanstrom, R., ... & Weeks, K. M. (2009). Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature, 460*(7256), 711–716.

Woese, C. R., Hinegardner, R. T., & Engelberg, J. (1964). Universality in the genetic code. *Science, 144*(3621), 1030-1031.

Woese, C. R. (1965). Order in the genetic code. *Proceedings of the National Academy of Sciences of the United States of America, 54*(1), 71.

Woese, C. R. (1973). Evolution of the genetic code. *Naturwissenschaften, 60*(10), 447-459.

Wong, J. T. (1975). A co-evolution theory of the genetic code. *Proceedings of the National Academy of Sciences of the United States of America, 72*(5), 1909.

Wuchty, S., Fontana, W., Hofacker, I. L., & Schuster, P. (1999). Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers, 49*(2), 145-165.

Yamao, F., Andachi, Y., Muto, A., Ikemura, T., & Osawa, S. (1991). Levels of tRNAs in bacterial cells as affected by amino acid usage in proteins. *Nucleic Acids Research., 19*(22), 6119-6122.

Yokobori, S. I., Suzuki, T., & Watanabe, K. (2001). Genetic code variations in mitochondria: tRNA as a major determinant of genetic code plasticity. *Journal of molecular evolution, 53*(4-5), 314-326.

Zia, R. K., Dong, J., & Schmittmann, B. (2011). Modeling translation in protein synthesis with TASEP: a tutorial and recent developments. *Journal of Statistical Physics, 144*(2), 405-428.

Zuker, M., & Stiegler, P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic acids research, 9*(1), 133-148.