On the evolution of codon usage bias

A Thesis Presented for

The Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Premal Shah

May 2011

Dedication

This dissertation is dedicated to my parents Rajesh and Heena for their constant support and encouragement.

Acknowledgements

I am greatly indebted to many folks who have directly and indirectly aided my growth both as a person and a scientist during my stay at the University of Tennessee. I would like to extend my sincere gratitude to my advisor Mike Gilchrist. He has influenced my approach to science more than anyone else. He is to be thanked for elevating my sense of caution and skepticism and rubbing off a bit of his quest for perfection in all he does. It would be an understatement to say that without Mike's push for focus I would have been lost by working on a myriad of things while completing nothing.

A special thanks to Jim Fordyce for infecting me with his insatiable passion for science. I have learnt a lot over beers with him than any class I have ever taken. He has been a constant source of inspiration both personally and scientifically.

I would also like to thank my committee members Lou Gross, Sergey Gavrilets and Russ Zaretzki for their support over the years. A special thanks to all the members of the GiGaOm and HOFF lab groups, especially Graham and Matt for introducing me to the slimy yet beautiful world of salamanders and snakes. I am also grateful for the generous financial support offered by the EEB department and NIMBioS, which not only helped support my research but also allowed me to present my research to wide audiences and make important personal connections. My stay in Knoxville couldn't have been made more enjoyable without the social and cultural support offered by Manthan (Indian Student Organization).

My philosophy on science and life has been greatly influenced by Arjun Krishnan and RS Prasanna. Arjun has been the perennial sounding board for all my ideas and immensely helpful in sorting out the good ones from the rest. His constant push for

justification of every last detail has been invaluable in bringing clarity to my own ideas.

At last, but more important than all, this work wouldn't have been possible without the unfailing love and affection of Samhita. She has been a pillar of strength during my bad times and has kept me grounded during the good ones. Her dedication to science and work has been greatly inspirational and influential.

"Our imagination is stretched to the utmost, not, as in fiction, to imagine things which are not really there, but just to comprehend those things which 'are' there."

-Richard Feynman

"When I don't understand something, I build a model"

-Michael Gilchrist

Abstract

The genetic code is redundant, with most amino acids coded by multiple codons. In many organisms, codon usage is biased towards particular codons. A variety of adaptive and non-adaptive explanations have been proposed to explain these patterns of codon usage bias. Using mechanistic models of protein translation and population genetics, I explore the relative importance of various evolutionary forces in shaping these patterns. This work challenges one of the fundamental assumptions made in over 30 years of research: codons with higher tRNA abundances leads to lower error rates. I show that observed patterns of codon usage are inconsistent with selection for translation accuracy. I also show that almost all the variation in patterns of codon usage in *S. cerevisiae* can be explained by a model taking into account the effects of mutational biases and selection for efficient ribosome usage. In addition, by sampling suboptimal mRNA secondary structures at various temperatures, I show that melting of ribosomal binding sites in a special class of mRNAs known as RNA thermometers is a more general phenomenon.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

One of the fundamental questions facing biologists is deciphering how the information in our genomes shapes our physiology and behavior. In the past, addressing this question has been difficult due to a lack of genomic data. However, with over 2000 genomes sequenced and the number expected to increase exponentially*, we are at the cusp of unraveling the intricacies of information contained in genomic sequences. This flood of data has also led to creation of entirely new fields of science including that of bioinformatics and systems biology. However, as Dobzhansky put it, "Nothing in biology makes sense except in the light of evolution" (Dobzhansky, 1973). Thus, my doctoral dissertation work is primarily based on explaining genomic patterns by combining models from both molecular and evolutionary biology. Specifically, this work integrates mechanistic models of specific biological processes such as protein translation with classical models in population genetics.

One of the earliest patterns to be discovered in genomic DNA was that of biases in codon usage (Fitch, 1976; Grantham *et al.*, 1980; Ikemura, 1981). The genetic code is highly redundant with multiple codons coding for a particular amino acid (Fig. 1.1). However, the frequency with which these codons are used within a genome are not uniform. There exists strong preference for certain codons over others. This preferential usage of codons is often referred to as Codon Usage Bias (CUB). Patterns of codon usage have been found in all three domains of life: Archaea, Eubacteria and

*http://www.ncbi.nlm.nih.gov/genomes/static/gpstat.html

Eukaryotes (Carbone *et al.*, 2003; Mougel *et al.*, 2004; Subramanian, 2008). Moreover, the codon usage changes not only among different organisms, but also between genes of a species as well as within a single gene. For many organisms this preferential use of certain codons is strongly correlated with corresponding tRNA abundances and gene expression levels (Ikemura, 1981; Dong *et al.*, 1996; Kanaya *et al.*, 1999). Identifying and explaining the evolutionary forces that shape these patterns has been the focus of a large number of studies spanning over three decades.

	U	C	A	G	
U	UUU F	UCU S	UAU Y	UGU C	U
	UUC F	UCC S	UAC Y	UGC C	C
	UUA L	UCA S	UAA X	UGA X	A
	UUG L	UCG S	UAG X	UGG W	G
C	CUU L	CCU P	CAU H	CGU R	U
	CUC L	CCC P	CAC H	CGC R	C
	CUA L	CCA P	CAA Q	CGA R	A
	CUG L	CCG P	CAG Q	CGG R	G
A	AUU I	ACU T	AAU N	AGU S	U
	AUC I	ACC T	AAC N	AGC S	C
	AUA I	ACA T	AAA K	AGA R	A
	AUG M	ACG T	AAG K	AGG R	G
G	GUU V	GCU A	GAU D	GGU G	U
	GUC V	GCC A	GAC D	GGC G	C
	GUA V	GCA A	GAA E	GGA G	A
	GUG V	GCG A	GAG E	GGG G	G

Figure 1.1: The Genetic Code

A variety of explanations have been put forth to explain these patterns of CUB. These explanations can be broadly classified into adaptive and non-adaptive mechanisms. Non-adaptive mechanisms include genetic drift, and biased mutation and gene conversion. Adaptive explanations for CUB comprise of selection for translation efficiency, selection for translation accuracy, selection against nonsense

errors, selection against ribosomal interference, and selection for DNA packaging. In multicellular eukaryotes such as humans, the effective population sizes of the species are low and the efficacy of selection in maintaining optimal codons in gene sequences is expected to be weak (Chamary *et al.*, 2006). Hence, in these organisms patterns of codon usage are thought to be primarily driven by non-adaptive forces. In contrast, in genomes of prokaryotes and unicellular eukaryotes, owing to their large effective population sizes, CUB in highly expressed genes is thought to be a result of natural selection. However, the relative importance of various selective forces in shaping these patterns remains an area of intense debate as multiple combinations of these forces can lead to similar patterns of codon usage. I briefly describe the various adaptive mechanisms proposed to explain CUB below.

1.1 Patterns and Explanations for CUB

1.1.1 Role of translation errors

Protein production is the most energetically expensive metabolic process within a cell (Warner, 1999; Akashi and Gojobori, 2002). However, like all biological processes, protein translation is prone to errors. The biological importance of these translation errors and their impact on coding sequence evolution, especially the evolution of codon usage bias (CUB), depends on both their effects on protein function and their frequencies. Translation errors fall into two categories: nonsense errors and missense errors.

Nonsense errors have a number of different causes such as ribosome drop-off, improper translation of release factors, and frame shifts (Menninger, 1977; Kurland, 1992; Kurland and Gallant, 1996). In addition to the indirect cost of ribosome usage, nonsense errors impose a direct assembly cost to the cell. These costs are proportional to the length of the peptide at the time of the error (Bulmer, 1991; Kurland, 1992; Eyre-Walker, 1996; Gilchrist and Wagner, 2006). Although

costly to produce, the vast majority of these incomplete peptides are expected to have no utility for the cell (Kurland and Gallant, 1996).

Missense errors are primarily caused by competition between cognate and near-cognate tRNAs (Kramer and Farabaugh, 2007; Fluitt *et al.*, 2007) followed by errors during initial tRNA selection and proof-reading (Rodnina and Wintermeyer, 2001; Gromadski and Rodnina, 2004; Wintermeyer *et al.*, 2004; Zaher and Green, 2009). Missense errors can lead to inactive or non-functional proteins, protein aggregation, nonsense errors and in some cases even cell death (Cornut and Willson, 1991; Kurland and Gallant, 1996; Zhao *et al.*, 2005; Lee *et al.*, 2006). However, unlike most nonsense errors which result in a non-functional protein, current data suggests that only ∼10-50% of missense errors disrupt protein function (Markiewicz *et al.*, 1994; Guo *et al.*, 2004).

Direct estimates of nonsense and missense error rates in prokaryotes suggest they occur with similar frequencies, i.e. on the order of 10^{-4} to 10^{-3} per codon (Manley, 1978; Tsung *et al.*, 1989; Jørgensen and Kurland, 1990; Ogle and Ramakrishnan, 2005; Kramer and Farabaugh, 2007)). Although these error rates may seem low, it is important to remember that these values are on a per codon basis and most coding sequences consist of hundreds of codons. For example, a translational error rate of $10^{-3.5}$ implies that for the average length protein ∼1 out of every 5 proteins will contain at least one error.

For over 30 years, the standard model of translation errors has implicitly assumed that for any given amino acid, the translation error rates are lowest for the codon with the highest tRNA abundances (Ikemura, 1981; Varenne *et al.*, 1984; Kramer and Farabaugh, 2007). Surprisingly, this assumption has not been adequately tested either theoretically or empirically until now. In Chapter 2 (Shah and Gilchrist, 2010b) we directly test this assumption and find that tRNA abundances are highly correlated, i.e., tRNAs with similar abundances are clustered within the genetic code. This pattern is observed across a wide range of bacterial genomes. Using a model of tRNA competition we also show that codons with higher tRNA abundances do

not always lead to lower error rates. If correct, this represents a major shift in our understanding of how tRNA abundances affect error rates and brings into question one of the fundamental assumptions made in decades of studies on codon usage patterns.

1.1.2 Role of translation efficiency

In addition to the cost of errors during protein translation, there are major indirect costs such as the cost of ribosome production. For example, in *S. cerevisiae* during log-growth phase, 2×10^3 ribosome are produced every minute tying up ∼60% of the cell's transcriptional machinery (Warner, 1999). Given their substantial cost, the efficient usage of these ribosomes during protein production is clearly advantageous and, therefore, one of the main explanations for the evolution of CUB (Bulmer, 1991).

In Chapter 3 we test the ability of a mechanistic model based on overhead cost of ribosome usage in protein production to explain and predict patterns of CUB. This is in contrast to most commonly used indices of CUB, such as F_{op} (Ikemura, 1981), *CAI* (Sharp and Li, 1986), and *CBI* (Bennetzen and Hall, 1982), which are both heuristic and aggregate measures of CUB and fail to explicitly define the factors responsible for the evolution of CUB. I find that our model can explain ∼92% of the observed variation in CUB across the *S. cerevisiae* genome indicating that cost of ribosomal usage may indeed be a dominant force in shaping CUB. Although, ours is not the first attempt at using mechanistic models to explain CUB in a population genetics context (Bulmer, 1991; Gilchrist, 2007), it is unique in its ability to estimate codon-specific parameters and quantitatively predict how codon frequencies change with gene expression. In addition the framework created in this study will allow explicit comparisons of various hypothesis proposed to explain patterns of codon usage and resolution of this long-standing debate. Moreover, the generality of our approach allows us to apply our model to any sequenced organism with available gene expression datasets.

1.1.3 Role of mRNA secondary structure in affecting gene expression

The protein production rate of a gene determines the efficacy of natural selection in affecting patterns of codon usage. Since protein translation is limited by the rate of translation initiation (Bulmer, 1991; de Smit and van Duin, 1990), selection for efficient usage of ribosomes would not only favor faster codons to increase the pool of free ribosomes within the cell but also affect the secondary structure of an mRNA for rapid initiation. This is due to the fact that secondary structure of an mRNA affects the rate at which ribosome 'jump' onto the mRNA. If the mRNA structure is such that the ribosome binding site (RBS) is sequestered in a closed hairpin structure, the ribosome cannot recognize it and hence cannot initiate protein translation (Yuzawa *et al.*, 1993; Nakahigashi *et al.*, 1995; Morita *et al.*, 1999; Narberhaus *et al.*, 2006). Hence, one would expect selection for less stable secondary structures near the RBS of an mRNA. It has been recently shown that mutations affecting the stability of mRNA secondary structures near the RBS site are correlated with changes in gene expression such that mRNAs with mutations that destabilize the structure lead to higher expression (Kudla *et al.*, 2009; Tuller *et al.*, 2010).

In Chapter 4 we test the relationship between mRNA secondary structure and temperature in RNA thermometers. RNA thermometers are genes whose expression level changes with temperature due to changes in the stability of its mRNAs (Yuzawa *et al.*, 1993; Nakahigashi *et al.*, 1995; Morita *et al.*, 1999). At lower temperatures, the sequence adopts a secondary structure that sequesters RBS of a gene, hence interfering with translation initiation by the ribosome. At higher temperatures, the mRNA melts, increasing the accessibility of the RBS leading to an increase in the initiation of translation and, in turn, its protein production rate (de Smit and van Duin, 1990; Yuzawa *et al.*, 1993; Chowdhury *et al.*, 2003; Narberhaus *et al.*, 2006). In order to test whether this 'melting' behavior is unique to RNA thermometers, we computationally sampled the distribution of the RNA structures

at various temperatures using Vienna - an RNA folding software. Although, known thermometers showed a higher rate of melting at their RBS compared to non-thermometers, contrary to our expectations these higher rates were not significant. I also did not find any significant differences between RNA thermometers from a range of γ-proteobacteria and *E. coli* non-thermometers. Although, in this study we did not link the effects of mRNA stability on patterns of codon usage, the methodology developed here would allow us to map such relationships explicitly.

Chapter 2

Effect of correlated tRNA abundances on translation errors and evolution of codon usage bias.

This chapter is a lightly revised version of a paper by the same name published in PLoS Genetics and co-authored with Michael A. Gilchrist.

Shah and Gilchrist. Effect of Correlated tRNA Abundances on Translation Errors and Evolution of Codon Usage Bias. PLoS Genet (2010) vol. 6 (9).

Abstract

Despite the fact that tRNA abundances are thought to play a major role in determining translation error rates, their distribution across the genetic code and the resulting implications have received little attention. In general, studies of codon usage bias (CUB) assume that codons with higher tRNA abundance have lower missense error rates. Using a model of protein translation based on tRNA competition and intra-ribosomal kinetics, we show that this assumption can be violated when tRNA abundances are positively correlated across the genetic code. Examining the distribution of tRNA abundances across 73 bacterial genomes from 20 different genera, we find a consistent positive correlation between tRNA abundances across the genetic code. This work challenges one of the fundamental assumptions made in over 30 years of research on CUB that codons with higher tRNA abundances have lower missense error rates and that missense errors are the primary selective force responsible for CUB.

2.1 Introduction

Protein production is the most energetically expensive metabolic process within a cell (Lobley *et al.*, 1980; Pannevis and Houlihan, 1992; Warner, 1999; Akashi and Gojobori, 2002). However, like all biological processes, protein translation is prone to errors. The biological importance of these translation errors and their impact on coding sequence evolution, especially the evolution of codon usage bias (CUB), depends on both their effects on protein function and their frequencies. Translation errors fall into two categories: nonsense errors and missense errors. Nonsense errors, also referred to as processivity errors, occur when a ribosome prematurely terminates translating a coding sequence. Missense errors occur when the wrong amino acid is incorporated into a growing peptide chain. Although many possible factors such as mRNA stability and recombination likely contribute to the evolution of CUB, selection against translation errors and biased mutation are thought to be the primary forces (Sharp and Li, 1986; Bulmer, 1991; Berg and Kurland, 1997; Kanaya *et al.*, 1999; Rocha, 2004; Drummond and Wilke, 2009; Gilchrist *et al.*, 2009).

Most researchers believe that CUB results primarily from selection against missense errors or, equivalently, for translational accuracy (see (Akashi, 1994, 2001; Arava *et al.*, 2005; Stoletzki and Eyre-Walker, 2007; Drummond and Wilke, 2009)). In addition to limited empirical observations, the main evidence cited as supporting this belief includes the fact that preferred synonymous codons (i.e. the codons over-represented in high expression genes) have higher cognate tRNA abundances and that these codons are also favored at evolutionarily conserved sites (Akashi, 1994, 2001). While the preferred codons may indeed be 'optimal' in some limited sense, as we demonstrate below, the idea that they minimize missense error rates is based on an overly simplistic understanding of the relationship between tRNA abundances and missense error rates.

The effect of missense errors on protein function is equivalent to a non-synonymous point mutation. Because amino acids with similar properties are clustered within the

genetic code (Grantham, 1974; Freeland and Hurst, 1998; Freeland *et al.*, 2000; Higgs, 2009), the genetic code is generally considered to be adapted to minimize the *phenotypic effects* of point mutations and missense errors. However, despite its importance, the adaptedness of tRNA abundances across the genetic code to reduce the *rate* of translation errors has received almost no attention. For instance, in *E. coli* the average nonsense and missense error rates are estimated to be on the order of 10^{-4} to 10^{-3} per codon, respectively (Andersson *et al.*, 1982; Bouadloun *et al.*, 1983; Precup and Parker, 1987; Kurland and Ehrenberg, 1987; Jørgensen and Kurland, 1990; Kramer and Farabaugh, 2007; Drummond and Wilke, 2009). This implies that for an average length gene of $\sim$ 300 amino acids, about 3-26% of its protein products will contain at least one translation error. However, since the only available estimates of missense error rates are for specific amino acid misincorporations (Andersson *et al.*, 1982; Bouadloun *et al.*, 1983; Precup and Parker, 1987), these rates are likely gross underestimates as they do not take into account all possible amino acid misincorporations at that codon.

Currently, missense errors are thought to be the result of competition between tRNAs with the right amino acid (cognates) and the ones with the wrong amino acids (near-cognates) for the codon at the ribosomal *A*-site (Varenne *et al.*, 1984; Gromadski and Rodnina, 2004; Kramer and Farabaugh, 2007). A near-cognate tRNA is characterized by a single codon-anticodon nucleotide mismatch and codes for an amino acid different from that of the *A*-site codon (Ogle *et al.*, 2001; Fluitt *et al.*, 2007; Zaher and Green, 2009). As a result of this competition, the rate of missense errors at a codon should be strongly affected by the abundances of both cognate and near-cognate tRNAs (Kramer and Farabaugh, 2007). For example, an increase in cognate tRNA abundances is predicted to lead to a decrease in a codon's missense error rate. In contrast, an increase in near-cognate tRNA abundances is predicted to lead to an increase in a codon's missense error rate (Kramer and Farabaugh, 2007).

Previous studies of CUB have generally assumed that amongst a set of synonymous codons, the one with the correspondingly highest tRNA abundance is the one with the lowest missense error rate. However, because missense error rates are thought to be a function of *both* cognate and near-cognate tRNA abundances, if tRNA abundances are positively correlated across the genetic code this assumption may not hold. In this study we ask a fundamental question, "Are tRNA abundances correlated across the genetic code?" Finding that tRNA abundances are indeed generally positively correlated across a wide range of prokaryotes, we then ask, "How does the distribution of tRNA abundances affect the relationship between codon translation and error rates?" This question is of critical importance because the currently favored explanation of CUB, what we will refer to as the standard model, implicitly assumes that codons with the highest translation rates are also the ones with the lowest missense error rates. Our results indicate that this basic assumption only holds for a limited subset of amino acids. As a result, our work strongly suggests that missense errors play a smaller role in the evolution of CUB than currently believed and that the observed patterns of codon conservation observed by Akashi and others are likely due to other selective forces such as selection for translational efficiency or against nonsense errors.

2.2 Results

We began our analysis by first assuming that the abundance of a tRNA species within a cell is proportional to its gene copy number (GCN). This relationship between tRNA abundance and GCN is often made in studies of CUB and has been observed in both prokaryotes and eukaryotes (Dong *et al.*, 1996; Kanaya *et al.*, 1999; Cognat *et al.*, 2008). We obtained GCNs of each tRNA type within an organism from the Genomic tRNA Database GtRNAdb (Chan and Lowe, 2009) for 73 bacterial genomes representing 50 species from 20 genera (see Table 2.1 for list of genomes analyzed).

Table 2.1: List of Genomes Analyzed

Aeromonas hydrophila ATCC 7966	Aeromonas salmonicida A449
Alkaliphilus metalliredigens QYMF	Alkaliphilus oremlandii OhILAs
Bacillus amyloliquefaciens FZB42	Bacillus anthracis Ames
Bacillus cereus ATCC14579	Bacillus cereus ATCC 10987
Bacillus cereus ZK	Bacillus cereus cytotoxis NVH 391-98
Bacillus subtilis	Bacillus thuringiensis Al Hakam
Bacillus thuringiensis konkukian	Bacillus weihenstephanensis KBAB4
Chromobacterium violaceum	Clostridium beijerinckii NCIMB 8052
Clostridium difficile 630	Clostridium perfringens
Clostridium perfringens ATCC 13124	Colwellia psychrerythraea 34H
Escherichia coli APEC O1	Escherichia coli CFT073
Escherichia coli C ATCC 8739	Escherichia coli E24377A
Escherichia coli HS	Escherichia coli K 12 substr DH10B
Escherichia coli K 12 substr W3110	Escherichia coli O157H7
Escherichia coli O157H7 EDL933	Escherichia coli SMS 3 5
Escherichia coli UTI89	Geobacillus kaustophilus HTA426
Geobacillus thermodenitrificans NG80-2	Heliobacterium modesticaldum Ice1
Klebsiella pneumoniae MGH 78578	Lactobacillus delbrueckii bulgaricus
Photobacterium profundum SS9	Lactobacillus delbrueckii bulgaricus BAA365
Pseudoalteromonas haloplanktis TAC125	Psychromonas ingrahamii 37
Salmonella typhimurium LT2	Shewanella ANA-3
Shewanella MR-4	Shewanella MR-7
Shewanella W3-18-1	Shewanella amazonensis SB2B
Shewanella baltica OS155	Shewanella baltica OS185
Shewanella baltica OS195	Shewanella denitrificans OS217
Shewanella frigidimarina NCIMB 400	Shewanella halifaxensis HAW EB4
Shewanella loihica PV-4	Shewanella oneidensis
Shewanella pealeana ATCC 700345	Shewanella putrefaciens CN-32
Shewanella sediminis HAW-EB3	Shewanella woodyi ATCC 51908
Shigella boydii CDC 3083 94	Shigella boydii Sb227
Shigella flexneri 2a	Shigella flexneri 2a 2457T
Shigella flexneri 5 8401	Shigella sonnei Ss046
Symbiobacterium thermophilum IAM14863	Vibrio cholerae
Vibrio cholerae O395	Vibrio fischeri ES114
Vibrio harveyi ATCC BAA-1116	Vibrio parahaemolyticus
Vibrio vulnificus CMCP6	Vibrio vulnificus YJ016
Yersinia pseudotuberculosis IP 31758	

We classified each amino acid based on its level of degeneracy i, where i represents the number of synonymous codons of that amino acid. As a result, each amino acid is placed in one of five different degenerate categories D_i ($i \in \{1, 2, 3, 4, 6\}$). For instance, alanine belongs to the D_4 class, while lysine belongs to the D_2 class as these amino acids are coded by 4 and 2 codons, respectively. Serine represents a special case as it is encoded by two disjoint degenerate subsets. As a result we treated each of these subsets as a separate amino acid. We calculated the correlation between GCN of a focal tRNA t_F and the sum of GCNs of neighboring tRNAs that coded for a different amino acid and differed from the focal tRNA's anticodon by a single base-pair, t_N (Table 2.2).

Table 2.2: List of Symbols

t_F	tRNA gene copy number of a focal codon
t_N	tRNA gene copy number of focal codon's neighbors
D_i	Set of amino acids with i synonymous codons
ρ_t	Correlation coefficient between t_F and t_N
ε_M	Missense error rate
ε_N	Nonsense error rate
R_c	Cognate elongation rate
R_n	Near-cognate elongation rate
R_d	Ribosomal drop-off rate
p_c	Probability of elongation by cognate tRNA per tRNA entry
p_n	Probability of elongation by near-cognate tRNA per tRNA entry
p_p	Probability of elongation by pseudo-cognate tRNA per tRNA entry
w	Wobble parameter

Figure 2.1 shows the distribution of correlation coefficients ρ_t between t_F and t_N for three degenerate classes of amino acids D_i within each of the genomes we examined.

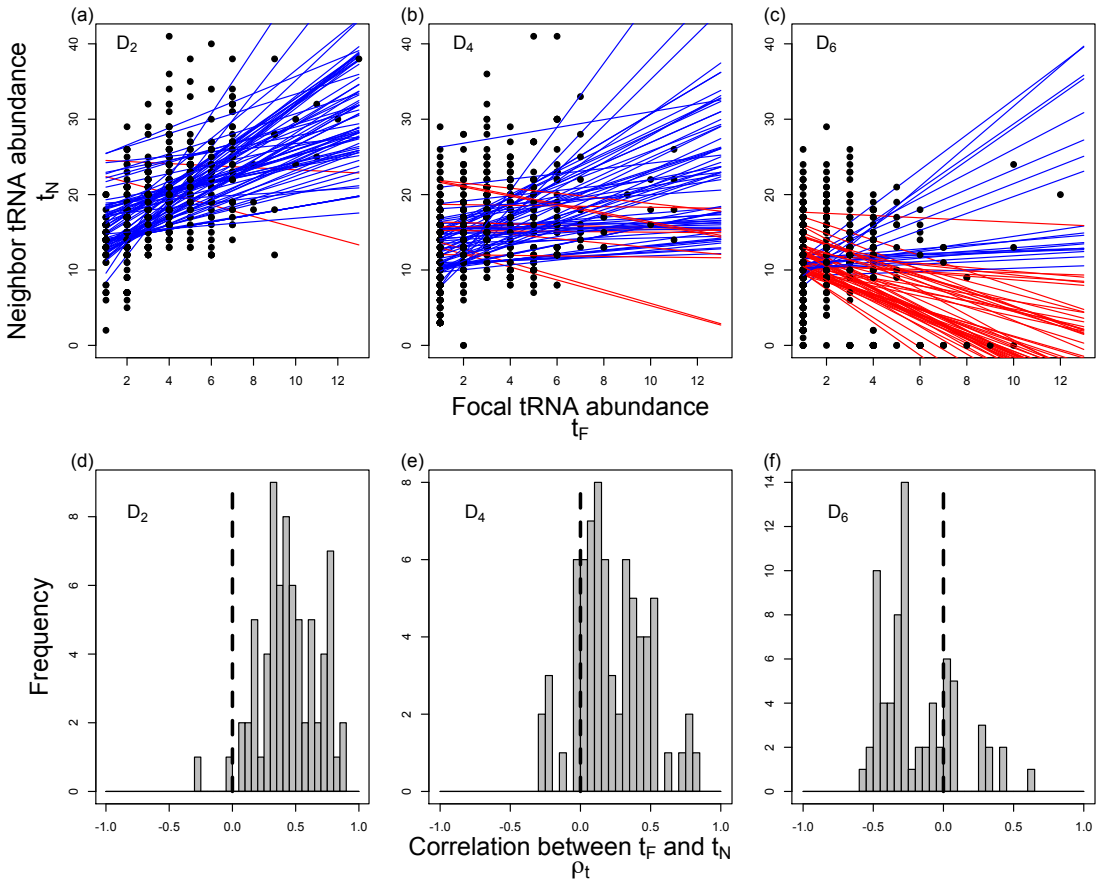


Figure 2.1: Correlation between a focal tRNA's abundance t_F and the abundance of its neighbors t_N, ρ_t across 73 prokaryotic genomes.

Each point in panels (A - C) represents a tRNA species that encodes an amino acid with degeneracy $D_i\,(i = \{2, 4, 6\})$. The solid lines represent the regression lines between t_F and t_N for each genome. Genomes with a negative ρ_t are coded in red, while genomes with a positive ρ_t are represented by blue lines. Panels (D - F) present the distribution of correlation coefficients ρ_t between t_F and t_N across all the genomes. The mean of the distribution of ρ_t values for all the three degenerate classes differ significantly from 0 (Wilcox test, $p < 10^{-7}$).

We find that the vast majority of genomes (69 out of 73 or $\sim 95\%$) show a positive relationship between the abundance of a focal tRNA species t_F and its one-step non-synonymous neighbors t_N, ρ_t (Binomial test, $p < 10^{-15}$, Figure 2.2).

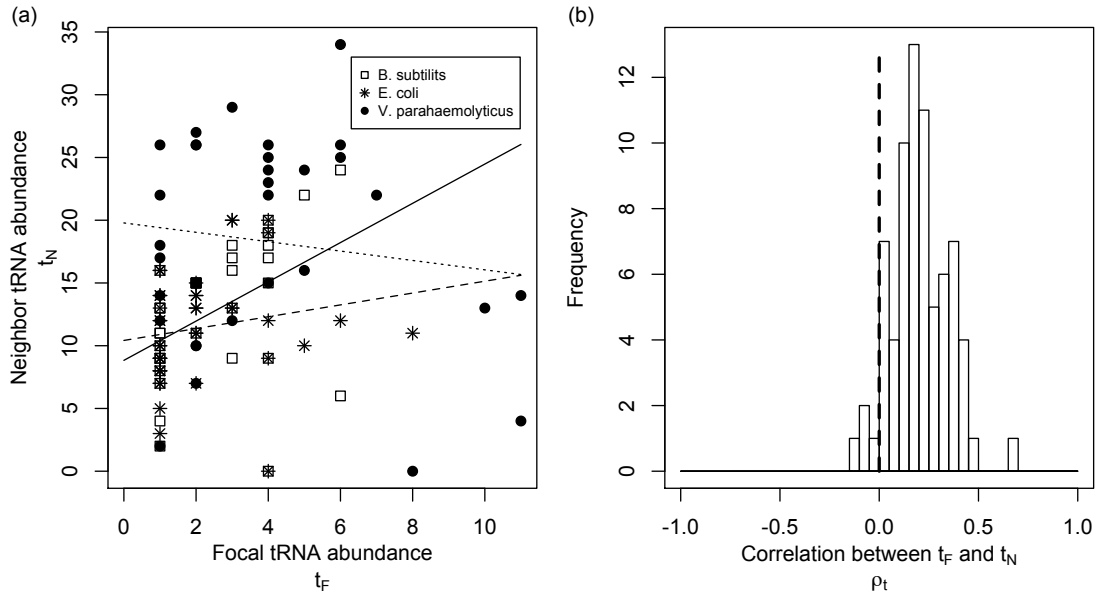


Figure 2.2: Correlation between a focal tRNA's abundance t_F and the abundance of its neighbors t_N across prokaryotic genomes.
Panel (a) represents the correlation between t_F and t_N across all amino acids for *B. subtilis*, *E. coli* and *V. parahaemolyticus* . Regression line between t_F and t_N for *B. subtilis*, *E. coli* and *V. parahaemolyticus* are represented by solid, dashed and dotted lines, respectively. Panel (b) shows the distribution of correlation coefficients ρ_t between t_F and t_N across 73 prokaryotic genomes. About 69 out of 73 genomes (Binomial test, $p < 10^{-15}$) have a positive relationship between t_F and t_N.

This indicates that tRNAs with similar abundances are closer to each other in the genetic code than expected under the implicit assumptions of the standard model. In other words, according to the standard model the tRNA abundances within the genetic code are predicted to be uncorrelated and the distributions of correlation

coefficients ρ_t in Figures 2.1 (d)-(f) are expected to be centered around 0. However, we find that under each of the degenerate classes of amino acids, D_2, D_4 and D_6, the distribution of ρ_t is significantly different from 0 (Wilcox test, $p < 10^{-7}$ for all D_i). Interestingly, we also find that the distribution of ρ_t differs considerably between degenerate classes of amino acids. tRNAs corresponding to amino acids in both D_2 and D_4 degenerate classes show a significant bias towards a positive correlation between t_F and t_N, whereas tRNAs in D_6 degenerate class are biased towards a negative correlation.

Since the frequency of amino acid usage within a genome is highly correlated with tRNA gene copy number (e.g. in *E. coli* $\rho = 0.632$, $p < 0.003$), the observed correlations may be the indirect result of amino acid usage bias. In addition to amino acid usage biases, the stereochemistry of codon-anticodon interactions forbids the existence of certain tRNA types (Lim and Curran, 2001), potentially contributing to the observed positive correlation among tRNA abundances. In order to address these inherent constraints on the distribution of tRNAs within the genetic code, we randomly distributed tRNA gene copies taking into account the stereochemical constraints, both with and without biased amino acid usage (see Figures 2.3 and 2.4).

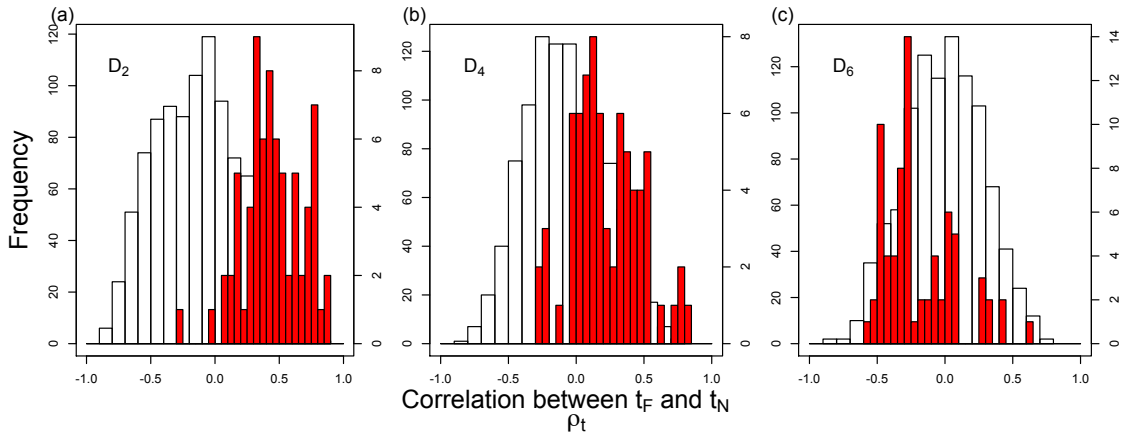


Figure 2.3: The distribution of correlation coefficients between a focal tRNA's abundance t_F and the abundance of its neighbors t_N, ρ_t.

Open bars represents the null distribution of ρ_t when tRNAs are randomly distributed across the genetic code, taking into account stereochemical constraints on possible tRNA anticodon types. Red bars represent the observed distribution of ρ_t across all 73 prokaryotic genomes. The observed distribution is significantly different from the null distribution (Kolmogorov-Smirnov test $p < 0.001$) across all three degenerate classes.

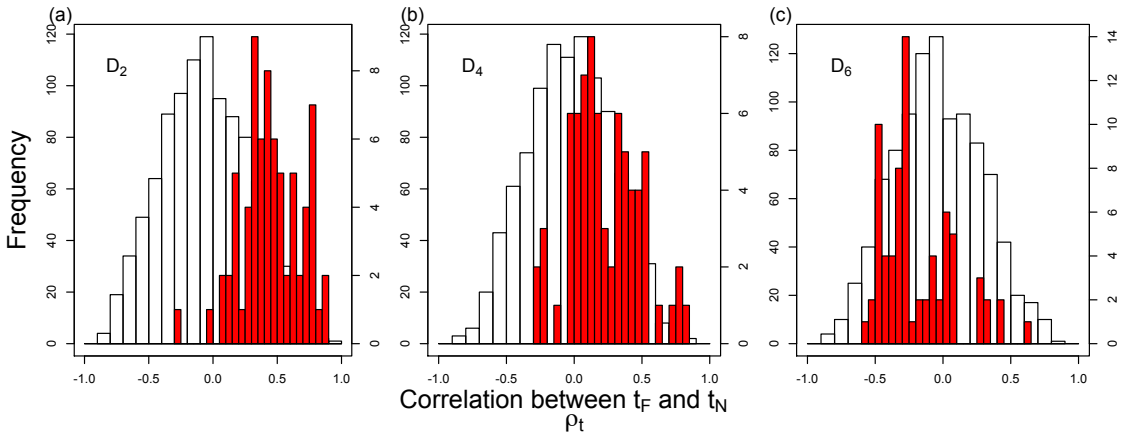


Figure 2.4: The distribution of correlation coefficients between a focal tRNA's abundance t_F and the abundance of its neighbors t_N, ρ_t.
Open bars represents the null distribution of ρ_t when tRNAs are randomly distributed across the genetic code prop, taking into account stereochemical constraints on possible tRNA anticodon types as well as the observed amino acid frequency distribution in *E. coli* genome. Red bars represent the observed distribution of ρ_t across all 73 prokaryotic genomes. The observed distribution is significantly different from the null distribution (Kolmogorov-Smirnov test $p < 0.001$) across all three degenerate classes.

We find that the observed distribution of ρ_t is significantly different from this more complex null distribution for all of the degenerate classes (Kolmogorov-Smirnov test $p < 0.001$ for all cases).

The distribution of tRNAs within the genetic code have important consequences with respect to translation errors and bias in codon usage. Codons with higher tRNA abundances than their coding synonyms are often referred to as 'optimal' codons (Drummond and Wilke, 2009) assuming they lead to fewer translation errors (Ikemura, 1985; Akashi, 1994; Kramer and Farabaugh, 2007). In light of the above results, we now ask the question, "Given that tRNA abundances are positively correlated in the genetic code, do higher cognate tRNA abundances always lead to fewer translation errors?"

2.2.1 Modeling translation errors

Following (Fluitt *et al.*, 2007), our model of translation errors takes into account competition between cognate and near-cognate tRNAs for the ribosomal *A*-site during translation. We also consider the kinetics of tRNA selection within a ribosome (Gromadski and Rodnina, 2004) and the effect of codon-anticodon wobble on these kinetics (Curran and Yarus, 1989). During protein translation, when a ribosome waits at a given codon, one of three outcomes is likely to occur: (a) elongation by cognate tRNA, (b) elongation by a near-cognate tRNA leading to a missense error or (c) spontaneous ribosomal drop-off, frameshift or recognition by release factors, any of which will lead to a nonsense error (Figure 2.5). The relative frequency of each of these outcomes determines the rates of missense and nonsense errors at a particular codon.

Assuming an exponential waiting process for a tRNA at codon i, the codon specific missense and nonsense error rates, ε_M and ε_N respectively, can be calculated as follows,

$$\varepsilon_M(i) = \frac{R_{\mathrm{n}}(i)}{R_{\mathrm{c}}(i) + R_{\mathrm{n}}(i) + R_{\mathrm{d}}} \tag{2.1}$$

$$\varepsilon_N(i) = \frac{R_{\mathrm{d}}}{R_{\mathrm{c}}(i) + R_{\mathrm{n}}(i) + R_{\mathrm{d}}} \tag{2.2}$$

where $R_c(i)$ is the codon specific cognate elongation rate, $R_n(i)$ is the codon specific near-cognate elongation rate, and R_d represents the background nonsense error rate (see *Methods* for details).

Using Equations (1) and (2), we calculated codon-specific missense and nonsense error rates for each bacterial genome. In order to understand the effect of codon degeneracy on the relationship between error rates and codon elongation rates, we categorized amino acids based on the number of their synonymous codons D_i as before. Given our model was parametrized from data on *E. coli*, we also checked for

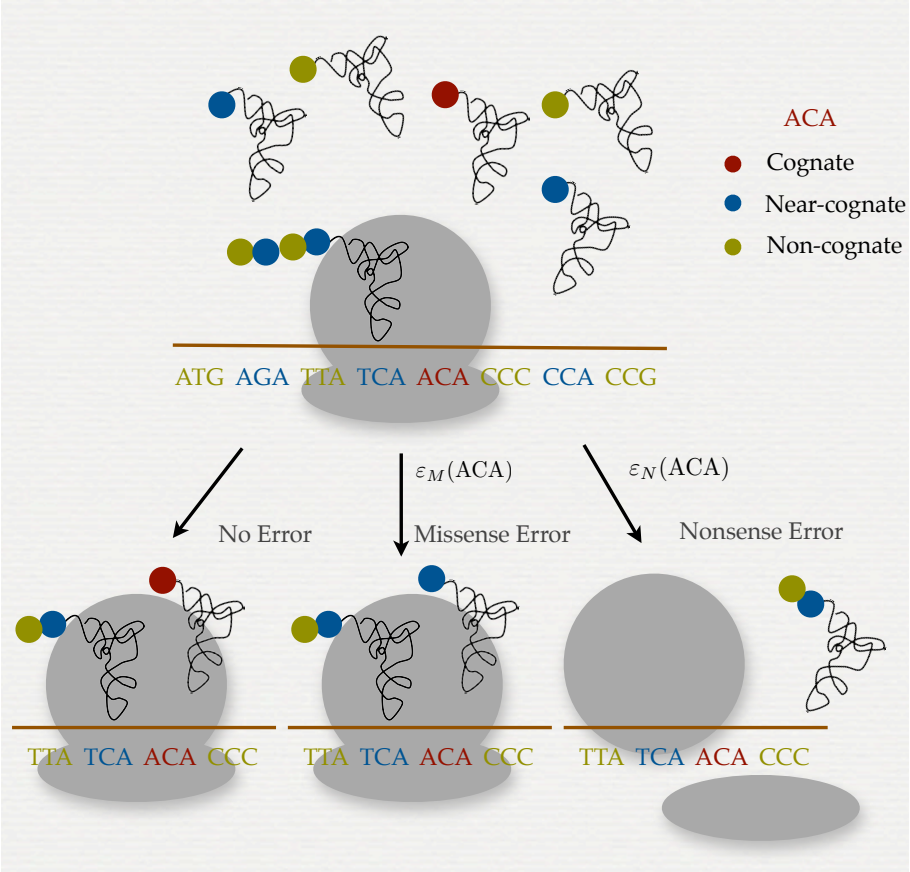


Figure 2.5: Model of translation errors.
During translation, a ribosome pauses at a codon (ACA in this case) waiting for a cognate tRNA. During this pause, one of the three processes can take place: elongation by cognate tRNAs leading to no translation error, elongation by a near-cognate tRNA leading to a missense error with rate ε_M or premature termination of translation due to recognition by release factors, spontaneous ribosome drop-off or frameshifting leading to a nonsense error with a rate ε_N.

the sensitivity of our analysis to changes in these parameters when extending it to other prokaryotes (Section 2.6.1).

2.2.2 Error Rates vs. Elongation Rates

Using *E. coli* strain K12/DH10B (K12) as an example, our estimates of codon-specific missense error rates ε_M ranged from $0 - 9.38 \times 10^{-3}$ with a median of 2.50×10^{-3}.

Six of the 61 sense codons have a predicted missense error rate of 0 as these codons have no near-cognate tRNA species (Table 2.3).

Table 2.3: List of codon-specific tRNAs, elongation rates and error rates in *E. coli*

AA	Codon	Cognates	Pseudo-cognates	Near-cognates	R_c	R_n	ε_M	ε_N
A	GCA	UGC	GGC	UCC, UAC, UGA, UGU, UGG, UUC	21.496	5.50E-02	2.55E-03	1.46E-04
A	GCC	GGC, UGC		GAC, GGU, GUC, GCC, GGA, GGG	27.218	5.86E-02	2.15E-03	1.15E-04
A	GCG	UGC	GGC	CGA, CGG, CCC, CGU	13.760	2.11E-02	1.53E-03	2.28E-04
A	GCU	GGC, UGC			22.061	0.00E+00	0.00E+00	1.43E-04
C	UGC	GCA		GCU, GUA, GCC, GGA, GAA, CCA	7.163	5.42E-02	7.51E-03	4.36E-04
C	UGU	GCA		ACG, CCA	4.584	2.16E-02	4.69E-03	6.83E-04
D	GAC	GUC		GGC, GAC, GUA, GUU, GCC, GUG, UUC	21.488	8.43E-02	3.91E-03	1.46E-04
D	GAU	GUC		UUC	13.752	1.64E-02	1.19E-03	2.28E-04
E	GAA	UUC		UGC, UCC, UAC, UUG, GUC, UUU	28.650	8.40E-02	2.92E-03	1.09E-04
E	GAG	UUC		GUC, CCC, CUG	18.336	2.49E-02	1.36E-03	1.71E-04
F	UUC	GAA		GAG, GAC, GCA, GAU, GUA, GGA, UAA, CAA	14.325	5.83E-02	4.05E-03	2.19E-04
F	UUU	GAA		UAA, CAA	9.168	8.21E-03	8.94E-04	3.43E-04
G	GGA	UCC	GCC, CCC	UGC, UAC, UCU, UUC	7.183	5.47E-02	7.56E-03	4.34E-04
G	GGC	GCC, UCC	CCC	GGC, GAC, GCU, GCA, GUC	32.952	3.81E-02	1.15E-03	9.53E-05
G	GGG	CCC, UCC	GCC	CCU, CCG, CCA	11.763	1.26E-02	1.07E-03	2.67E-04
G	GGU	GCC, UCC	CCC	ACG	22.638	1.64E-02	7.25E-04	1.39E-04
H	CAC	GUG		GAG, GUA, GUU, UUG, GUC, GGG, CUG	7.163	6.65E-02	9.20E-03	4.35E-04
H	CAU	GUG		ACG, UUG, CUG	4.584	3.39E-02	7.34E-03	6.81E-04
I	AUA	CAU	GAU	UAG, UAC, UGU, UUU, UCU, UAA, CAU	36.685	9.83E-02	2.67E-03	8.55E-05
I	AUC	GAU	CAU	GAG, GAC, GCU, GUU, GGU, GAA, CAU	21.521	8.32E-02	3.85E-03	1.46E-04
I	AUU	GAU		CAU	13.785	3.28E-02	2.38E-03	2.28E-04
K	AAA	UUU		GUU, UGU, UUG, UCU, UUC	42.976	5.06E-02	1.18E-03	7.31E-05
K	AAG	UUU		GUU, CCU, CAU, CUG, CGU	27.504	7.01E-02	2.54E-03	1.14E-04
L	CUA	UAG	GAG, CAG, UAA	UAC, UUG, UGG	7.189	3.31E-02	4.58E-03	4.35E-04
L	CUC	GAG, UAG	CAG	GAC, GAU, GAA, GUG, GGG	11.477	3.78E-02	3.28E-03	2.73E-04
L	CUG	CAG, UAG	GAG, CAA	CGG, CCG, CAU, CUG	33.243	4.95E-02	1.49E-03	9.45E-05
L	CUU	GAG, UAG	CAG	ACG	8.898	1.64E-02	1.84E-03	3.53E-04
L	UUA	UAA	UAG, CAA	UAC, UGA, GAA	7.171	3.31E-02	4.59E-03	4.36E-04
L	UUG	CAA, UAA	CAG	CGA, GAA, CAU, CCA	11.764	4.95E-02	4.19E-03	2.66E-04
M	AUG	CAU		GAU, CAG, CCU, CAA, CGU	57.301	4.57E-02	7.97E-04	5.49E-05

Table 2.3: (continued)

AA	Codon	Cognates	Pseudo-cognates	Near-cognates	R_c	R_n	ε_M	ε_N
N	AAC	GUU		GCU, GAU, GUA, GGU, GUC, UUU, GUG	28.650	7.91E-02	2.75E-03	1.09E-04
N	AAU	GUU		UUU	18.336	2.46E-02	1.34E-03	1.71E-04
P	CCA	UGG	CGG, GGG	UAG, UGC, UGA, UGU, UUG	7.171	3.34E-02	4.63E-03	4.36E-04
P	CCC	GGG, UGG	CGG	GAG, GGC, GGU, GGA, GUG	11.464	3.37E-02	2.93E-03	2.74E-04
P	CCG	CGG, UGG	GGG	CGA, CAG, CCG, CUG, CGU	11.751	4.24E-02	3.60E-03	2.67E-04
P	CCU	GGG, UGG	CGG	ACG	8.886	1.64E-02	1.84E-03	3.53E-04
Q	CAA	UUG	CUG	UAG, UGG, UUU, GUG, UUC	14.334	5.34E-02	3.71E-03	2.19E-04
Q	CAG	CUG, UUG		CGG, CAG, GUG, CCG	23.493	2.90E-02	1.23E-03	1.34E-04
R	AGA	UCU	CCU	UCC, GCU, UGU, UUU	7.167	3.89E-02	5.39E-03	4.36E-04
R	AGG	CCU, UCU	CCG	GCU, CAU, CC, CCA, CGU	11.751	5.36E-02	4.54E-03	2.66E-04
R	CGA	ACG	UCU, CCG	UAG, UCC, UUG, UGG	17.199	2.11E-02	1.22E-03	1.83E-04
R	CGC	ACG	CCG	GAG, GCU, GCA, GCC, GUG, GGG	18.340	3.75E-02	2.04E-03	1.71E-04
R	CGG	CCG, ACG	CCU	CGG, CAG, CCC, CCA, CUG	24.357	3.78E-02	1.55E-03	1.29E-04
R	CGU	ACG	CCG		28.655	0.00E+00	0.00E+00	1.10E-04
S	UCA	UGA	CGA, GGA	UGC, UGU, UGG, UAA	7.175	2.52E-02	3.50E-03	4.37E-04
S	UCC	GGA, UGA	CGA	GGC, GCA, GUA, GGU, GAA, GGG	18.627	4.60E-02	2.46E-03	1.68E-04
S	UCG	CGA, UGA	GGA	CGG, CCA, CAA, CGU	11.755	2.11E-02	1.79E-03	2.67E-04
S	UCU	GGA, UGA	CGA		13.470	0.00E+00	0.00E+00	2.33E-04
U	ACA	UGU	GGU, CGU	UGC, UGA, UGG, UUU, UCU	7.180	5.01E-02	6.93E-03	4.35E-04
U	ACC	GGU, UGU	CGU	GGC, GCU, GAU, GUU, GGA, GGG	18.631	5.47E-02	2.93E-03	1.68E-04
U	ACG	CGU, UGU	GGU	CGA, CGG, CCU, CAU	18.917	4.74E-02	2.50E-03	1.66E-04
U	ACU	GGU, UGU	CGU		13.474	0.00E+00	0.00E+00	2.33E-04
V	GUA	UAC	GAC	UAG, UGC, UCC, UUC, UAA	35.821	4.19E-02	1.17E-03	8.77E-05
V	GUC	GAC, UAC		GAG, GGC, GAU, GUC, GCC, GAA	35.813	6.29E-02	1.75E-03	8.77E-05
V	GUG	UAC	GAC	CAG, CAU, CCC, CAA	22.929	5.97E-02	2.60E-03	1.37E-04
V	GUU	GAC, UAC			30.656	0.00E+00	0.00E+00	1.03E-04
W	UGG	CCA		GCA, CGA, CCU, CCG, CCC, CAA	7.163	2.49E-02	3.46E-03	4.37E-04
Y	UAC	GUA		GCA, GUU, GUC, GGA, GAA, GUG	21.488	5.39E-02	2.50E-03	1.46E-04
Y	UAU	GUA			13.752	0.00E+00	0.00E+00	2.29E-04
Z	AGC	GCU		GCA, GAU, GUU, GGU, GCC, CCU, UCU	7.163	6.79E-02	9.38E-03	4.35E-04
Z	AGU	GCU		ACG, CCU, UCU	4.584	2.46E-02	5.34E-03	6.82E-04

These rates are higher than recent empirical estimates of missense error rates in *E. coli*, which vary from $2.0 \times 10^{-4} - 3.6 \times 10^{-3}$ with a median value of 3.4×10^{-4} (Kramer and Farabaugh, 2007). This is likely due to the fact that the missense error estimates in (Kramer and Farabaugh, 2007) were for specific amino acid misincorporations, whereas, the values predicted here indicate the rate of all possible missense errors at a given codon. Our predicted rates of codon-specific nonsense errors ε_N in *E. coli* ranged from $5.49 \times 10^{-5} - 6.83 \times 10^{-4}$ with a median of 2.19×10^{-4} (Table 2.3).

We find that on average both missense ε_M and nonsense error rates ε_N decrease with an increase in cognate elongation rates R_c (Figure 2.6).

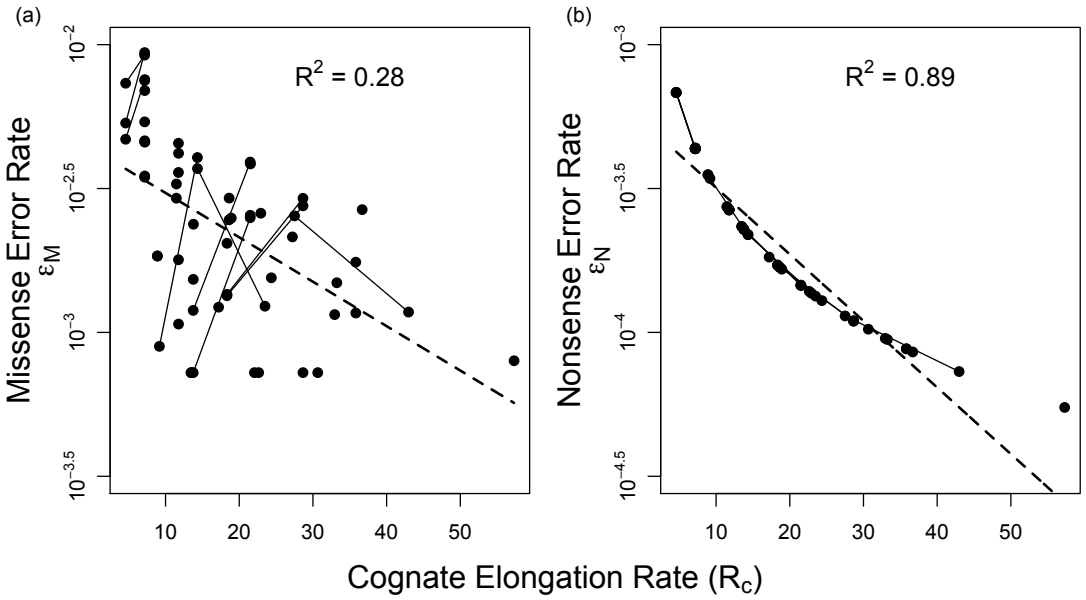


Figure 2.6: Correlation of translation error rates ε with cognate elongation rate R_c in *E. coli*.

We find that rates of both (A) missense ε_M and (B) nonsense errors ε_N are negatively correlated with the rate of elongation by cognate tRNAs at that codon. The dashed line indicates the regression line between R_c and ε. This is consistent with expectations under the standard model. However, in the case of twofold degenerate amino acids (D_2), whose two codons are joined together by solid lines, we see that ε_M increases with R_c for 8 out of 10 amino acids. In the case of ε_N every amino acid showed a decrease in ε_N with R_c.

These results seem, on first glance, largely consistent with the standard model for inferring translation errors from tRNA abundances, which assumes that ε decreases with R_c. However, because R_n varies between synonymous codons, for about half of the amino acids (10 out of 21) ε_M is actually greater for the codon with the highest R_c value. This holds even when empirical estimates of tRNA abundances in *E. coli* (Dong *et al.*, 1996) are used instead of tRNA gene copy numbers (see Figure 2.7).

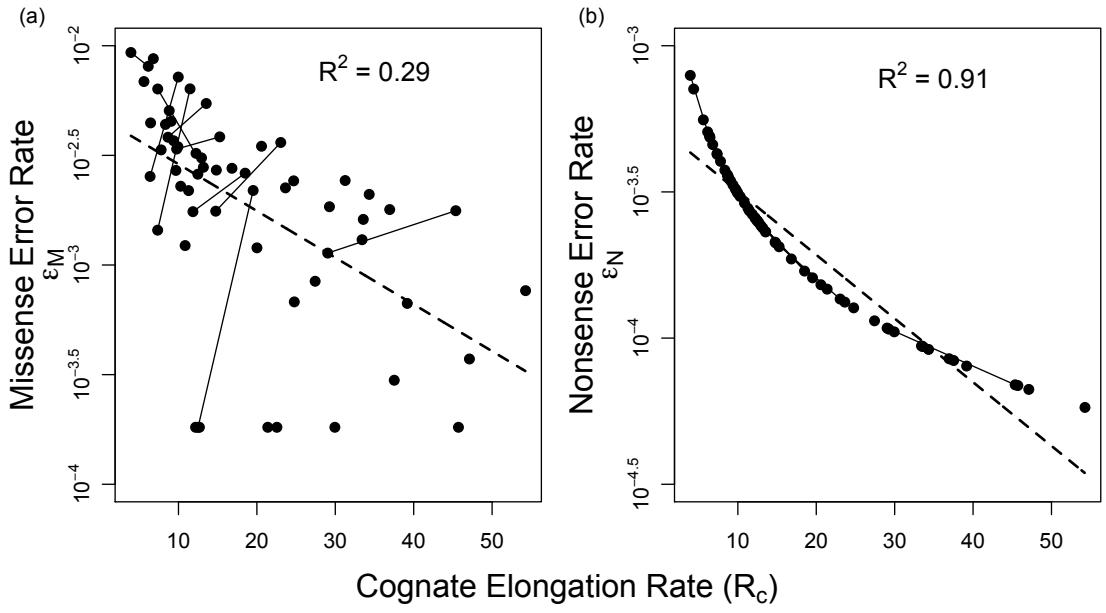


Figure 2.7: Correlation of translation error rates ε with cognate elongation rate R_c using empirical estimate of tRNA abundances.
We find that rates of both (a.) missense ε_M and (b.) nonsense errors ε_N are negatively correlated with the rate of elongation by cognate tRNAs at that codon. The dashed line indicates the regression line between R_c and ε. These results are consistent with the results obtained using tRNA gene copy numbers as proxies for tRNA abundances.

This result is *inconsistent* with expectations under the standard model that implicitly assumes a codon-independent rate of elongation by near-cognate tRNAs, R_n. If the abundance of a focal tRNA t_F and its neighbors t_N are uncorrelated, then the only factor that affects ε_M is R_c. However, as shown earlier, t_F and t_N are positively correlated (Figure 2.1). Thus, the estimates of ε_M of synonymous codons of an amino acid depend not only on their individual R_c but also on the slope of the relationship between R_c and R_n. If the rate of increase of R_n with R_c is higher than the relative increase in R_c, then codons with higher cognate elongation rates R_c are expected to have *higher* missense error rates ε_M (Figure 2.8).

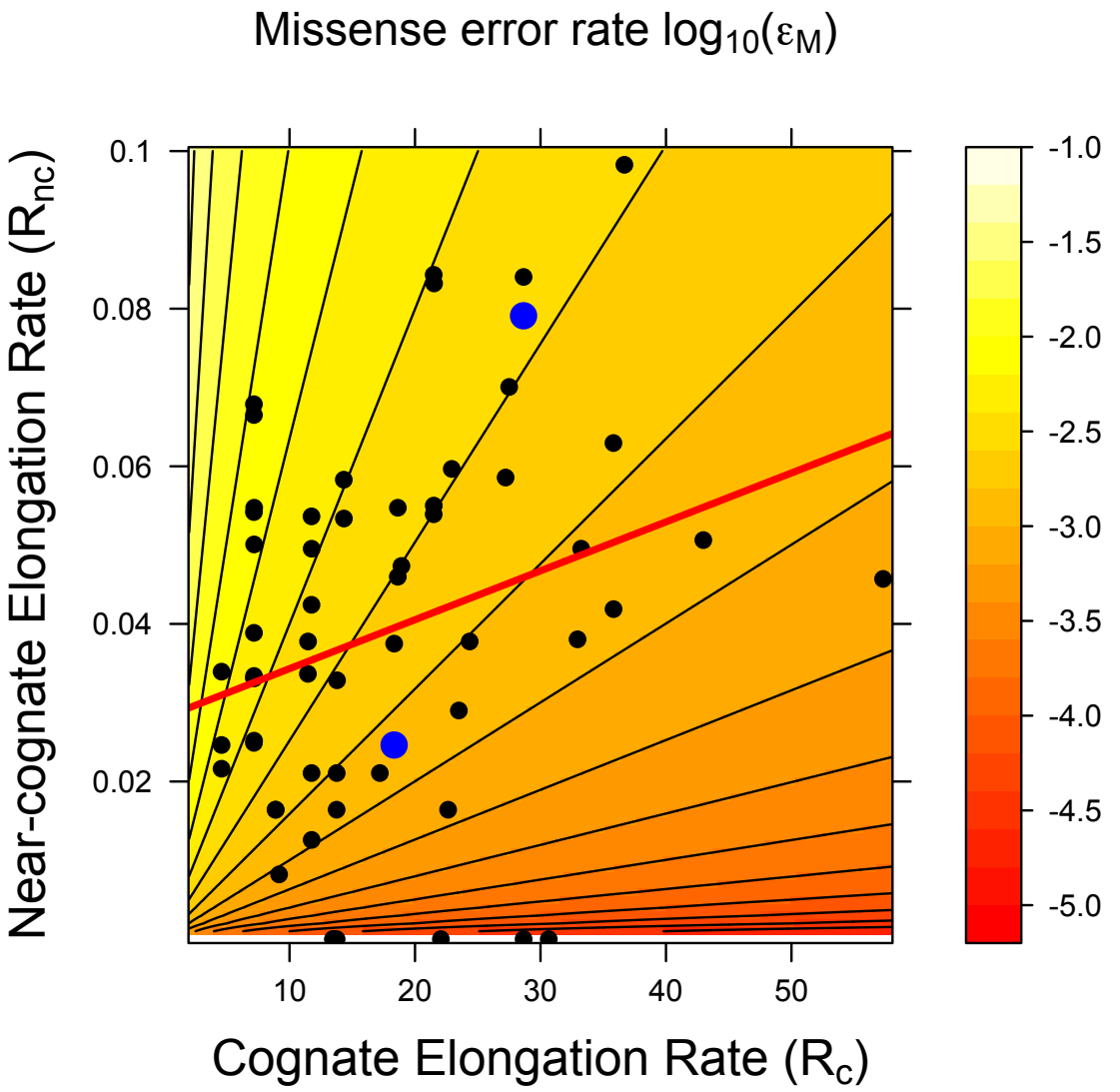


Figure 2.8: Contour plot of missense error rates $\log_{10}(\varepsilon_M)$ with cognate R_c and near-cognate R_n elongation rates.

The black dots represent $\log_{10}(\varepsilon_M)$ of codons in *E. coli*. Blue dots are the two codons of amino acid asparagine (N). In the case of asparagine, the codon with a higher R_c has a higher ε_M as it also has a much higher R_n. The regression line between observed R_c and R_n in *E. coli* is represented as a solid red line. The positive correlation between R_c and R_n, explains why codons with higher R_c sometimes have a higher missense error rate.

Interestingly, 8 out of the 10 D_2 amino acids in *E. coli* K12 showed a positive relationship between R_c and ε_M. Specifically, we would expect ε_M to increase with R_c whenever the condition $\frac{dR_n}{dR_c} > \frac{R_n}{R_c}$ is satisfied. Thus, among the synonymous codons of an amino acid in *E. coli*, the codon with the lowest ε_M is often not the codon with the highest R_c. This points to a fundamental change in our understanding of the relationship between tRNA abundances and missense errors and which codons minimize their occurrence.

Interestingly, these results are also consistent with the limited empirical estimates of codon-specific missense error rates. For instance, (Precup and Parker, 1987) used *E. coli* to estimate rates at which the asparagine codons AAC and AAU were mistranslated by $\text{tRNA}^{\text{Lys}}_{\text{UUU}}$. As expected, the authors found that the AAC codon, with a higher R_c had a lower rate of mistranslation by $\text{tRNA}^{\text{Lys}}_{\text{UUU}}$ than AAU, with a lower R_c. Our model makes the same prediction when considering this specific subset of missense errors. However, when considering the overall missense error rates at AAC and AAU codons due to tRNA^{Lys}, tRNA^{Ser}, tRNA^{Thr}, tRNA^{Asp}, tRNA^{His}, tRNA^{Tyr} and tRNA^{Ile}(all one-step neighbors), we come to a very different prediction. Specifically we find that even though AAC has a higher R_c than AAU, it also has a much higher R_n rate. As a result, the *overall* missense error rate for AAC is actually predicted to be higher than AAU. This result illustrates how focusing on only a subset of possible missense errors at a codon, as all previous experiments have done, provides an incomplete and potentially misleading picture.

In contrast to missense error rates, our model predicts ε_N will consistently decline with an increase in R_c, suggesting that nonsense errors may be playing a larger role in driving CUB than commonly accepted (Arava *et al.*, 2005).

2.2.3 Intra- and Inter-specific Variation in the Relationship between Elongation and Error Rates

In order to evaluate the relationship between cognate elongation rate, R_c, and error rates, we looked across 73 bacterial genomes for inter-specific variation and 11 strains of *E. coli* for intra-specific variation. As before, we categorized amino acids based on the degeneracy of their synonymous codons for each genome. We calculated the fraction of amino acids within each category that showed a *negative* relationship between R_c and error rates, ε_M and ε_N (Figure 2.9) as expected under the standard model where the abundances of tRNAs are assumed to be uncorrelated.

For both intra- and inter-specific datasets we find that synonymous codons with a higher R_c have a lower nonsense error rate ε_N for all amino acids, irrespective of the degenerate class D_i they belong to. However, in the case of missense errors, the relationship between R_c and ε_M depends on the amino acid degeneracy D_i as previously observed in *E. coli* K12 (Figure 2.6). Amino acids with two synonymous codons (D_2) show a strong bias towards a *positive* relationship between R_c and ε_M, both intra- and inter-specifically (Binomial test, $p = 1.5 \times 10^{-10}$ and $p < 2.2 \times 10^{-16}$, respectively). In the case of isoleucine, the only amino acid in D_3, there exists no bias towards a positive or a negative relationship between cognate elongation and missense error rates (Binomial test, intra-specific $p = 0.548$ and interspecific $p = 0.349$). Interestingly 4-fold degenerate amino acids show a bimodal distribution of the fraction of genomes with a negative relationship, and the two 6-fold degenerate amino acids (arginine and leucine) show a strong bias towards negative correlation between R_c and ε_M (Binomial test, intra-specific $p = 4.7 \times 10^{-7}$ and interspecific $p < 2.2 \times 10^{-16}$). The differences in the relationship between ε_M and R_c across degenerate classes are similar to the differences in the correlation between t_F and t_N across these classes (Figure 2.1).

Although the patterns we observe are complex and vary with amino acid degenerate classes, the assumption underlying the standard model that higher cognate

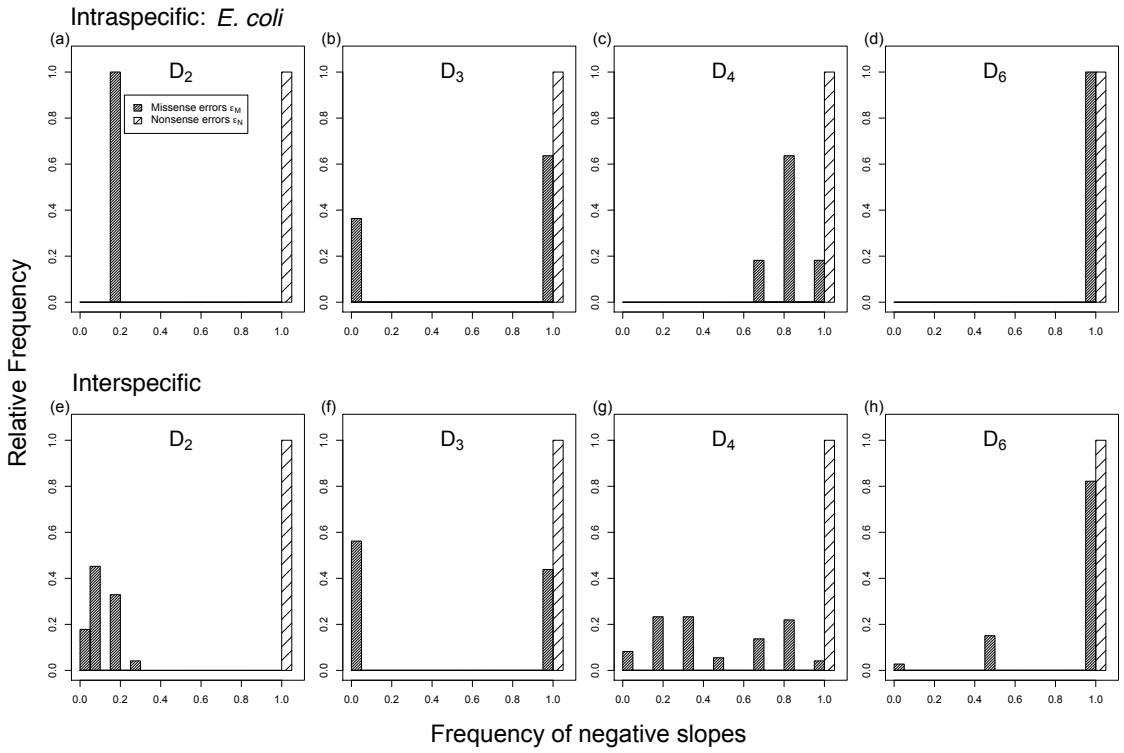


Figure 2.9: Frequencies of negative relationships between cognate elongation rate R_c and translation errors ε.
Panels (A - D) represent the distribution of *E. coli* strains that show amino acid specific negative relationship between R_c and ε, while panels (E - H) represent the distribution of 73 genomes for the same. Amino acids in every degenerate class (D_i) show a negative relationship between cognate elongation rate R_c and nonsense error rates (ε_N) both intra-specifically as well as inter-specifically. A majority of amino acids in the 2-fold degenerate class (D_2) show an increase in missense error rate ε_M with R_c across genomes. As the degeneracy of amino acids increases, we see an increase in the frequency of the expected negative relationship between ε_M and R_c across *E. coli* strains as well as other bacterial species.

tRNA abundance codons will have the lowest translation error rates is predicted to be clearly violated in the case of missense errors – a finding consistent both across bacterial genomes and across various *E. coli* strains. We also find that the positive relationship between missense error rates ε_M and R_c observed within certain amino acids is insensitive to moderate changes in parameter estimates of background nonsense error rates, and wobble parameters (Section 2.6.1).

2.3 Discussion

For over 30 years, the standard model of translation errors has implicitly assumed that for any given amino acid, the translation error rates are lowest for the codons with the highest tRNA abundances (Ikemura, 1981; Varenne *et al.*, 1984; Kramer and Farabaugh, 2007). With respect to missense errors ε_M, this prediction was based on the implicit and unstated assumption that the distribution of tRNA abundances across the genetic code are uncorrelated. Here we show a consistent positive correlation between the abundance of a tRNA and its one-step mutational neighbors across a wide array of prokaryotes. In order to understand the effects of this relationship on translation errors, we developed a simple model for estimating codon-specific error rates based on the distribution of tRNA gene copy number of a species. Our model takes into account tRNA competition, wobble effects, and intra-ribosomal kinetics of elongation to predict rates of missense and nonsense errors. To our knowledge, ours is the first model to integrate all these factors for estimating translation errors. Using our model, we find that on *average*, both missense and nonsense error rates of a codon decrease with an increase in its cognate tRNA elongation rate. This average behavior is consistent with expectations under the standard model of how codon specific error rates scale with cognate tRNA abundance (Akashi, 1994; Stoletzki and Eyre-Walker, 2007; Kramer and Farabaugh, 2007; Drummond and Wilke, 2008). However, the expected relationship between error rates and cognate tRNA abundances does not hold at finer scales of individual amino acids, the relevant scale for the evolution of CUB.

For about half of the amino acids (10 out of 21) in *E. coli* K12, synonymous codons that have higher cognate elongation rates R_c also have higher missense error rates ε_M. This counterintuitive behavior is due to the fact that tRNA abundances within the genetic code are positively correlated, which leads to an increase in ε_M with R_c, an important pattern that has been overlooked by previous researchers. We find a positive correlation between the abundance of a focal tRNA t_F and

that of its neighbors t_N in 69 out of 73 genomes examined here. In addition, the 4 genomes that show a negative ρ_t (*E. coli* O157H7, *E. coli* O157H7-EDL933, *Photobacterium profundum* SS9, *Vibrio parahaemolyticus*) also show evidence of a high degree of horizontal gene transfer. Interestingly we also find that the differences in the relationship between t_F and t_N across amino acid degenerate classes is mirrored in the correlation between ε_M and R_c. In contrast to ε_M, the nonsense error rates ε_N of synonymous codons decrease with an increase in R_c for every amino acid across every genome we analyzed. This is due to the fact that increasing either R_c or R_n leads to a decrease in ribosomal wait time at that codon which, in turn, leads to a lower ε_N. Thus with respect to ε_N, a positive correlation between tRNA abundances actually accentuates the advantage of using codons with higher tRNA abundances. These results lend further support to the hypothesis that nonsense errors play an important but under-appreciated role in the evolution of CUB (Gilchrist, 2007; Gilchrist *et al.*, 2009).

The role of tRNA competition has been recognized as an important factor in affecting translation error rates (Varenne *et al.*, 1984; Fluitt *et al.*, 2007; Kramer and Farabaugh, 2007). However, previous studies on the relationship between error rates and tRNA abundances have focused primarily on the effects of modifying cognate tRNA abundances and ignored the effects of near-cognate tRNA abundances. Consistent with our model behavior, (Kramer and Farabaugh, 2007) showed that when $\text{tRNA}^{\text{Arg}}_{\text{UCU}}$ was over-expressed, it led to a decrease in the missense error rate ε_M at codons for which the tRNA was a cognate: AGA and AGG. However, if a higher expression level of $\text{tRNA}^{\text{Arg}}_{\text{UCU}}$ reduces the frequency of ε_M at codons AGA and AGG, why is it not fixed in the population? We argue that increasing the abundance of a given tRNA may not always be adaptive. For instance, over-expressing $\text{tRNA}^{\text{Arg}}_{\text{UCU}}$ will also lead to an increase in ε_M at nearby non-synonymous codons - AAA, ACA, AUA, etc., a testable prediction not considered by (Kramer and Farabaugh, 2007). The trade-offs between reducing ε_M at one codon at the expense of increasing ε_M at nearby codons has not been explored. However, these trade-offs likely play an important role

in shaping the evolution of tRNA gene copy number and force us to reconsider the evolutionary causes of CUB.

Currently, many researchers believe that selection for translational accuracy, i.e., against missense errors, is a primary force driving the evolution of CUB (see (Akashi, 1994; Arava *et al.*, 2005; Drummond *et al.*, 2005; Stoletzki and Eyre-Walker, 2007)). This belief largely rests on the interpretation of two facts. Firstly, preferred codons are generally those with the highest corresponding tRNA abundances and secondly, sites that are highly conserved and thought to have large effects on protein structure and function, use preferred codons more often than their coding synonyms (Akashi, 1994). Selection for translational accuracy is usually tested using Akashi's test by identifying evolutionarily conserved sites in protein sequences and checking whether they are coded by preferred codons (Akashi, 1994; Drummond *et al.*, 2006; Stoletzki and Eyre-Walker, 2007; Drummond and Wilke, 2009). In light of the above results, we need to revisit the underlying assumptions of Akashi's test (Akashi, 1994). Although, our analysis predicts that a considerable number of amino acids have a positive relationship between missense error rates, ε_M and cognate elongation rates R_c, many amino acids in *E. coli* are still predicted to conform to the standard model of lower ε_M with higher R_c. Indeed, in the case of *Drosophila* species used in the original Akashi's paper (Akashi, 1994), only 4 out of 21 amino acids are predicted to have a positive relationship between ε_M and R_c. Thus, we argue that the relationship between ε_M and R_c are highly species and amino acid specific and that selection for translation accuracy cannot explain all of the observed CUB at conserved sites. In addition to selection for translational accuracy, selection against nonsense errors (Gilchrist and Wagner, 2006; Gilchrist, 2007; Gilchrist *et al.*, 2009), mRNA stability (Bulmer, 1991) and protein misfolding due to ribosome stalling (Kimchi-Sarfaty *et al.*, 2007; Tsai *et al.*, 2008) have been shown to affect CUB. In fact, recent evidence suggests that the speed of translating a codon also affects protein folding (Kimchi-Sarfaty *et al.*, 2007; Tsai *et al.*, 2008; Marin, 2008). The presence of a codon with a low R_c, increases the ribosomal waiting time at a codon

potentially leading to alternate protein folds. This directly affects the functionality and stability of the protein. Thus, a codon with a higher R_c at a conserved site, as observed by Akashi and others, could be under selection to prevent protein misfolding due to an entirely different mechanism unrelated to missense errors. Thus, we would like to stress that the definition of preferred codons used in the Akashis test is based on the genome-wide frequency of codon usage and not on any fundamental biological process. Although, we do not dispute the fact that certain codons are preferred over others at conserved sites, we simply point that the presence of these preferred codons at conserved sites cannot be explained entirely by selection against missense errors and that other selective forces must be responsible for the maintenance of these codons.

CUB often increases with gene expression, such that highly expressed genes tend to use codons with a higher cognate elongation rate R_c (Ikemura, 1985; Greenbaum *et al.*, 2003; Gilchrist *et al.*, 2009). Thus, these genes would have lower nonsense error rates and wait times, but not necessarily lower missense error rates. This might appear paradoxical, as the failure to minimize missense error rate would presumably increase the probability that a translated protein would be rendered nonfunctional and be selected against. However, the deleterious effects of a high missense error rate can be mitigated by an increased robustness of highly expressed genes. According to (Kellogg and Juliano, 1997; Drummond *et al.*, 2005; Wilke and Drummond, 2006), highly expressed genes are expected to evolve at a slower rate and also be extremely functionally robust to missense errors. If this is the case, then missense errors in highly expressed genes may not have much of an effect on protein function. These genes maybe perfectly poised for trading off an elevated missense error rate for faster elongation and fewer nonsense error rates.

When it comes to mitigating the effects of non-synonymous mutations and missense errors, the genetic code has been described as "one in a million" (Freeland and Hurst, 1998). This is due to the fact that amino acids with similar chemical properties are in a genetic 'neighborhood', thus reducing the phenotypic effect of any

point mutation or missense error. However, unlike point mutations, the frequency of missense errors depends on the distribution of tRNA within the genetic code. The distribution of tRNA abundances is usually attributed to the coevolution between codon usage and tRNA abundances (Wong, 1975; Ardell and Sella, 2001; Vetsigian and Goldenfeld, 2009). However, these studies have not taken into account how changes in tRNA abundances affect the rate of translation errors at neighboring codons. The degree to which the distribution of tRNA abundances within the genetic code is adapted to minimize translation errors remains largely unexplored. Our work suggests that understanding the trade-offs between missense and nonsense errors would provide significant insights into the evolution of tRNA abundances within the genetic code. We believe building mechanistic models of translation errors, as shown here, will help further our understanding of the evolution of tRNA abundances across the genetic code.

2.4 Methods

2.4.1 tRNA competition

Assuming an exponential waiting process and simple diffusion, the rates at which cognate and near-cognate tRNAs enter the ribosomal *A*-site will be proportional to their abundances. As a result, translation error rates of a codon will depend, in part, on the relative abundances of its cognate and near-cognate tRNAs (Kramer and Farabaugh, 2007). Following (Dong *et al.*, 1996; Kanaya *et al.*, 1999; Cognat *et al.*, 2008), we use the GCN of a tRNA as a proxy for its abundance.

2.4.2 Intra-ribosomal dynamics

Discrimination between cognate, near-cognate and non-cognate tRNAs takes place in the peptidyl transfer step of elongation. Since the underlying process is stochastic, there is a non-zero probability that when a cognate tRNA enters the *A*-site it will

be rejected or a near-cognate tRNA will be accepted (Gromadski and Rodnina, 2004). These probabilities are a function of the kinetic rate constants of various steps involved within the peptidyl transfer and translocation processes during tRNA elongation for both cognate and near-cognate tRNAs (Gromadski and Rodnina, 2004; Blanchard *et al.*, 2004b,a) (Section 2.6.2). Based on the rate constants for cognate and near-cognate tRNAs from (Gromadski and Rodnina, 2004) and equations from (Fluitt *et al.*, 2007), we estimated the probability of elongation of a codon by a cognate and near-cognate tRNA per tRNA entry into the ribosomal *A*-site to be $p_c = 6.52 \times 10^{-1}$ and $p_n = 6.2 \times 10^{-4}$, respectively (Section 2.6.2).

2.4.3 Wobble effects

One of the factors affecting the rate constants in the intra-ribosome kinetic model described above, is the effect of codon-anticodon wobble. (Gromadski and Rodnina, 2004) proposed that a wobble mismatch between a codon and its cognate tRNA anticodon, will affect its kinetic rate constants (Section 2.6.2) and consequently reduce the probability of elongation by that tRNA. Based on (Curran and Yarus, 1989; Lim and Curran, 2001), we assume that a purine-purine or pyrimidine-pyrimidine wobble reduces the probability of a cognate tRNA being accepted p_c, by 40%. This reduction in p_c is consistent with estimates based on the kinetic rate constants estimated by (Kothe and Rodnina, 2007) for Ala_{GCC} codon that is recognized by $\text{tRNA}^{\text{Ala}}_{\text{UGC}}$ through a pyrimidine-pyrimidine wobble. Similarly, based on (Curran and Yarus, 1989) ,we assume that a non-canonical purine-pyrimidine wobble (GU/AC) would reduce p_c by 36%.

In addition, some codons can be recognized by cognate tRNAs through a non-standard wobble as described by (Agris, 1991; Agris *et al.*, 2007). For instance, C-U and C-A anticodon-codon interactions are considered nonstandard owing to their stereochemistry and thermodynamic constraints. Hence, even though anticodon $\text{tRNA}^{\text{Ala}}_{\text{CGC}}$ does not lead to a missense error when translating the codon Ala_{GCU}, it

is considered nonstandard translation due to its C-U wobble. We call these tRNAs 'pseudo-cognates'. We assume that the probability of elongation of a codon by pseudo-cognates p_p is the same as that of near-cognate tRNAs, i.e., $p_p = p_n$.

2.4.4 Estimation of cognate and near-cognate elongation rates

In order to predict per codon missense and nonsense error rates, we calculated the rates of elongation by cognate and pseudo-cognate tRNAs vs. near-cognate tRNAs at each codon. The cognate elongation rate for codon i is given by

$$R_c(i) = a \left(\sum_{j \in \mathbb{S}_{\mathrm{c}}(i)} t_j p_{\mathrm{c}} w_{j,i} + \sum_{j \in \mathbb{S}_{\mathrm{p}}(i)} t_j p_{\mathrm{p}} w_{j,i} \right) \tag{2.3}$$

where $\mathbb{S}_{\mathrm{c}}(i)$ is the set of cognate tRNAs for codon i, $\mathbb{S}_{\mathrm{p}}(i)$ represents the set of pseudo-cognate tRNAs, t_j represents the gene copy number of j^{th} tRNA species, and $w_{j,i}$ is the reduction in elongation probability due to wobble mismatch.

Similarly, the rate at which near-cognate tRNAs elongate codon i is given by

$$R_n(i) = a \sum_{j \in \mathbb{S}_{\mathrm{n}}(i)} t_j p_{\mathrm{n}} w_{j,i} \tag{2.4}$$

where $\mathbb{S}_n(i)$ is the set of near-cognate tRNAs with respect to codon i. The parameter a represents a scaling constant between tRNA gene copy number GCN and elongation rate. For *E. coli*, we used a value of $a = 10.992\ s^{-1}$, so that the harmonic mean of elongation rates of all codons was $\overline{R_c + R_n} \sim 12.5\ aa/s$ (Andersson *et al.*, 1982; Varenne *et al.*, 1984; Sørensen *et al.*, 1989).

We assume that nonsense errors occur primarily due to spontaneous drop-off of ribosomes at a given codon when it is waiting for a tRNA. As a result, the nonsense error rate due to spontaneous ribosomal drop-off, $R_d(i)$, is codon independent and occurs at a constant rate. (Jørgensen and Kurland, 1990) measured a nonsense

error rate of 1 per 4000 codons. If we assume $\overline{R_c + R_n} \sim 12.5\ aa/sec$, then the background rate of nonsense errors is $R_d = 3.146 \times 10^{-3}\ s^{-1}$.

2.5 Acknowledgments

We would like to thank Arjun Krishnan and Justin Vaughn for providing helpful suggestions and comments on this manuscript. We also thank the editors and three anonymous reviewers for their constructive criticisms and suggestions that have greatly improved this manuscript.

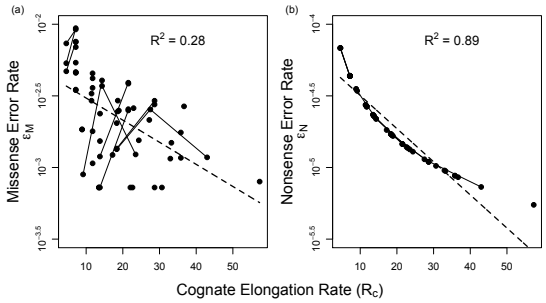
2.6 Supporting Information

2.6.1 Parameter Sensitivity

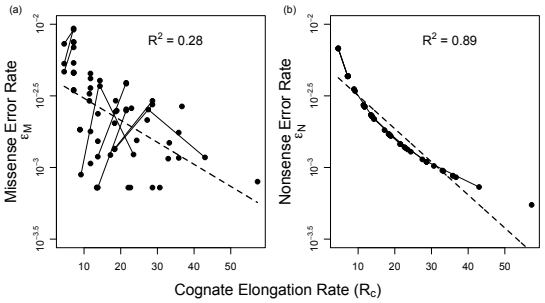
Since our model was parametrized using empirical data for *E. coli*, we checked for the sensitivity of our analyses to changes in underlying parameters. Specifically, we changed the wobble parameters (w_{RR} and w_{RY}) and the rate of premature termination (R_d). We checked for the sensitivity to parameters by visually comparing the correlation of error rates (ε_M and ε_N) versus cognate elongation rate (R_c) as well as by comparing the distribution of these correlations across amino acids both intra- and inter-specifically.

Cognate elongation rate versus error rates

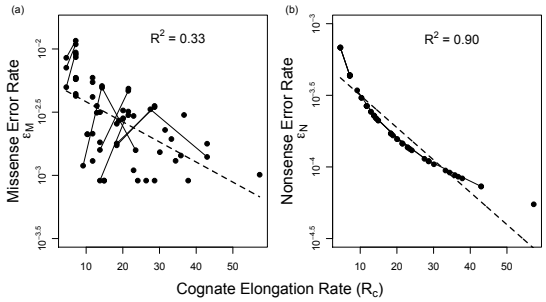
We find no qualitative difference in the relationship between cognate elongation and error rates when the rate of premature termination (R_d) was both increased and decreased by an order of magnitude. However, we did see a corresponding change in the overall nonsense error rate of codons, as expected.



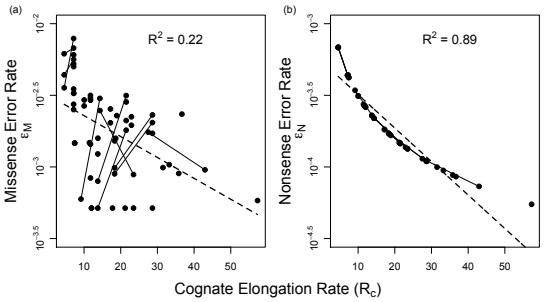
(a) $R_d = 3.146 \times 10^{-4}$



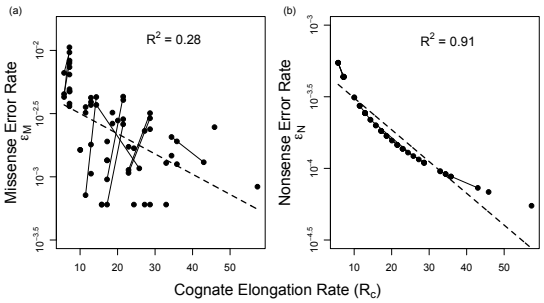
(b) $R_d = 3.146 \times 10^{-2}$



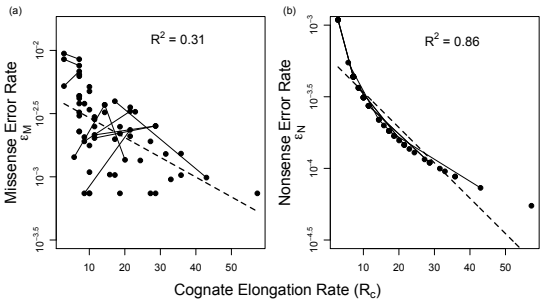
(c) $w_{RR} = 0.8, w_{RY} = 0.64$



(d) $w_{RR} = 0.4, w_{RY} = 0.64$



(e) $w_{RR} = 0.6, w_{RY} = 0.8$



(f) $w_{RR} = 0.6, w_{RY} = 0.4$

Figure 2.10: Sensitivity of model behavior to changes in parameters.

2.6.2 Estimating probability of elongation at a codon during one tRNA insertion attempt

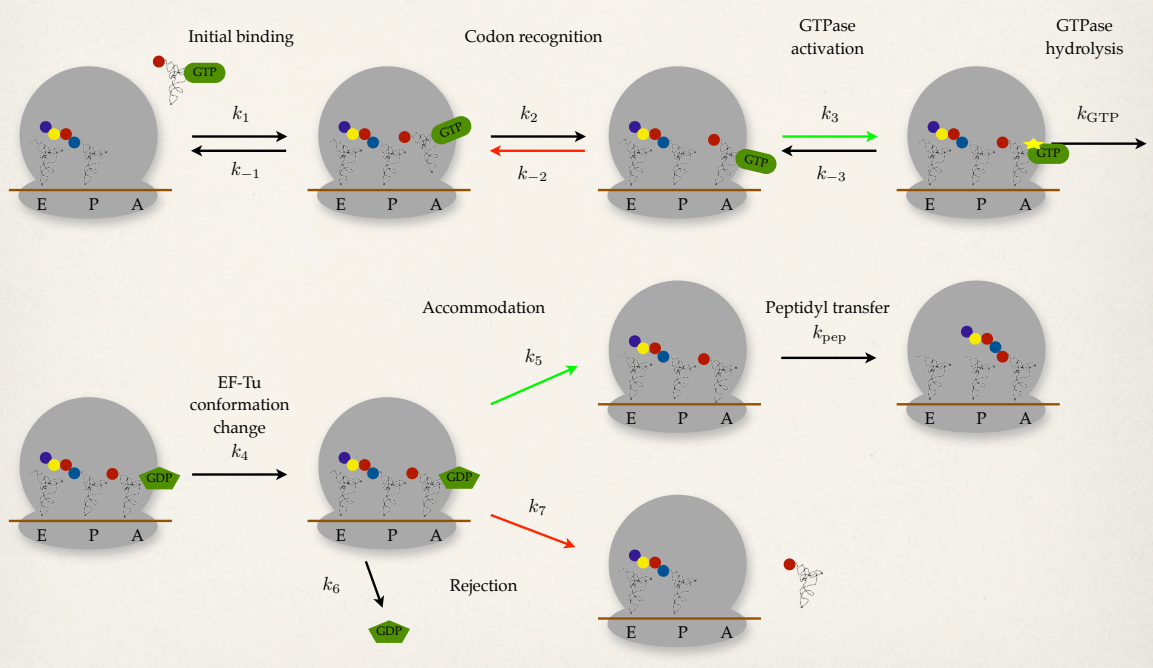


Figure 2.11: Kinetic model of tRNA selection
The kinetic model as adapted from Gromadski and Rodnina (2004)

Rate Constant	k_1 ($\mu m^{-1} s^{-1}$)	k_{-1} (s^{-1})	k_2 (s^{-1})	k_{-2} (s^{-1})	k_3 (s^{-1})	k_{GTP} (s^{-1})	k_4 (s^{-1})	k_5 (s^{-1})	k_7 (s^{-1})	k_{pep} (s^{-1})
Cognate	140	85	190	0.23	260	1000	1000	1000	60	200
Near-cognate	140	85	190	80	0.4	1000	1000	60	1000	200

Table 2.4: Rate constants for the kinetic model of tRNA selection

Using Eqn. (5) from Fluitt, et.al. (2007), we estimated the probability of elongation as

$$p = \frac{P_{23}P_{34}P_{67}}{P_{23}P_{34} + P_{21}} \tag{2.5}$$

$$P_{23} = \frac{k_2}{k_2 + k_{-1}} \qquad P_{34} = \frac{k_3}{k_3 + k_{-2}} \qquad P_{67} = \frac{k_5}{k_5 + k_7} \qquad P_{21} = \frac{k_{-1}}{k_{-1} + k_2} \tag{2.6}$$

Plugging in the values for cognate and near-cognate tRNAs, we find $p_c = 6.52 \times 10^{-1}$ and $p_n = 6.2 \times 10^{-4}$.

Chapter 3

Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and genetic drift

This chapter is a lightly revised version of a paper by the same name submitted in Proc. Natl. Acad. Sci. and co-authored with Michael A. Gilchrist.

Abstract

The genetic code is redundant with most amino acids using multiple codons. In many organisms, codon usage is biased towards particular codons. Understanding the adaptive and non-adaptive forces driving the evolution of codon usage bias (CUB) has been an area of intense focus and debate in the fields of molecular and evolutionary biology. However, their relative importance in shaping genomic patterns of CUB remains unsolved. Using a nested model of protein translation and population genetics, we show that observed gene level variation of CUB in *S. cerevisiae* can be explained almost entirely by selection for efficient ribosomal usage, genetic drift and biased mutation. The correlation between observed codon counts within individual genes and our model predictions is 0.96. Although a variety of factors shape patterns of CUB at the level of individual sites within genes, our results suggest that selection for efficient ribosome usage is a central force in shaping codon usage at the genomic scale. In addition, our model allows direct estimation of codon-specific mutation rates and elongation times and can be readily applied to any organism with high throughput expression datasets. More generally, we have developed a natural framework for integrating models of molecular processes to population genetics models to quantitatively estimate parameters underlying fundamental biological processes such as protein translation.

3.1 Introduction

For many organisms the preferential use of certain codons, commonly referred to as codon usage bias (CUB), is strongly correlated with corresponding tRNA abundances and expression levels (Ikemura, 1981; Dong *et al.*, 1996). Explanations for these correlations abound; the most favored ones include selection against translational errors (Akashi, 1994; Drummond and Wilke, 2009; Gilchrist, 2007), selection for translational efficiency (Bulmer, 1991; Akashi and Eyre-Walker, 1998; Coleman *et al.*, 2008), effects on protein folding (Kimchi-Sarfaty *et al.*, 2007), and stability of mRNA secondary structures (Kudla *et al.*, 2009; Tuller *et al.*, 2010). Since different combinations of these factors could lead to very similar patterns of codon usage, their relative importance in shaping evolution of CUB is unknown (Arava *et al.*, 2005; Kudla *et al.*, 2009; Shah and Gilchrist, 2010b). We believe that this uncertainty over their relative importance is, in large part, due to lack of mechanistic models of processes hypothesized to give rise to these patterns (for exceptions see (Bulmer, 1991; Gilchrist and Wagner, 2006; Shah and Gilchrist, 2010b)). While most theories of codon usage predict that the degree of bias in codon usage should increase with gene expression (Ikemura, 1981; Drummond and Wilke, 2009; Gilchrist *et al.*, 2009), they lack any specific quantitative predictions about the rate and nature of these changes. This is because most commonly used indices of CUB, such as F_{op} (Ikemura, 1981), *CAI* (Sharp and Li, 1986), and *CBI* (Bennetzen and Hall, 1982), are both heuristic and aggregate measures of CUB and fail to explicitly define the factors responsible for the evolution of CUB (for exceptions see (dos Reis *et al.*, 2004; Gilchrist *et al.*, 2009)). In contrast, we show that a mechanistic model of protein translation that explicitly includes the effects of biased mutation, genetic drift, and selection for efficient ribosome usage can explain the genome wide codon usage patterns in *S. cerevisiae*. Although, ours is not the first attempt at using mechanistic models to explain CUB in a population genetics context (Bulmer, 1991; Gilchrist, 2007), it is

unique in its ability to estimate codon-specific parameters and quantitatively predict how codon frequencies change with gene expression. We find that our model can explain ∼92% of the observed variation in CUB across the *Saccharomyces cerevisiae* genome.

3.2 Model

Protein synthesis is the most energetically expensive process within a cell (Wagner, 2005). During the log-phase of growth in *S. cerevisiae*, about 60% of transcriptional machinery is devoted to making about 2000 ribosomes every minute (Warner, 1999). Since ribosomes are large complexes with finite lifespan and are expensive to manufacture, one would expect strong selection for their efficient use during protein translation (Kurland, 1987; Bulmer, 1991; Lovmar and Ehrenberg, 2006; Hershberg and Petrov, 2008). Here we explicitly define selection for efficient use of ribosomes as selection for translational efficiency (Bulmer, 1991; Lovmar and Ehrenberg, 2006). Since codons that have longer elongation times tie up ribosomes on the mRNA leading to an inefficient usage, these codons should be selected against. Thus, in the absence of other factors, selection for translation efficiency should favor coding sequences that use codons with shorter elongation times and the strength of this selection should increase with gene expression (Bulmer, 1991; Akashi and Eyre-Walker, 1998; Akashi, 2003; Hershberg and Petrov, 2008). If selection for translational efficiency is a major force driving the evolution of CUB in *S. cerevisiae*, then we should be able to predict the CUB of a gene based on the differences in elongation times of synonymous codons, mutational bias, and its expression level.

We model the cost of protein production explicitly in terms of ATP usage as it is common currency for energy consumption within a cell (Alberts *et al.*, 2008). Based on the work in (Gilchrist, 2007; Gilchrist *et al.*, 2009), we begin our model by first noting that in the absence of translation errors, the expected cost for production

of a single protein is simply

$$\eta(\vec{x}) = C \sum_{i=1}^{61} x_i t_i, \tag{3.1}$$

where x_i is the number of codons of type i among the 61 sense codons used within a given coding sequence $\vec{x} = \{x_1, x_2, \ldots x_{61}\}$, t_i is the expected elongation time for codon i, and C is a scaling factor that represents the overhead cost of ribosome usage in ATP/sec. Codons that have shorter elongation times will lead to lower costs η, and hence, are expected to be selected over their coding synonyms. Based on the work in (Gilchrist, 2007; Gilchrist *et al.*, 2009) we assume an exponential fitness function $w(\vec{x}|\phi) \propto e^{-q\phi\eta(\vec{x})}$, where q is the scaling constant (sec/ATP) determining the relationship between the rate of ATP usage and fitness w and ϕ is a measure of gene expression, specifically protein production rate (proteins/sec). It is important to note the distinction between the protein production rate and the translation rate of a ribosome across an mRNA. This lack of distinction has been the source of confusion over the role of gene expression in shaping patterns of codon usage in the past (Bulmer, 1991; Plotkin and Kudla, 2011).

In addition, although protein production rate of a gene changes during a single cell's lifetime, the ϕ value used here is the target time-averaged rate at which the protein will be produced. In this scenario, a change from an optimal codon to a suboptimal codon does not affect ϕ but instead affects the cost of meeting the target ϕ. Using the cost of producing a protein η as the phenotype, we calculate the probability of observing a particular coding sequence given its expression level, $P(\vec{x}|\phi)$. $P(\vec{x}|\phi)$ is defined for each coding sequence in the synonymous codon genotype space S_c for a given protein. Under the Fisher-Wright process (Wright, 1969; Gavrilets, 2004; Sella and Hirsh, 2005) this probability is,

$$P\left(\vec{x}|\phi\right) \propto w\left(\vec{x}|\phi\right)^{N_e} \prod_{i=1}^{61} \mu_i^{x_i} \tag{3.2}$$

where N_e is the effective population size and μ_i is the sum of mutation rates to codon i from its synonymous codons (Sella and Hirsh, 2005). Simply put, $P(\vec{x}|\phi)$, the probability of observing a particular synonymous codon genotype for a given protein is a combined function of mutation bias $\prod_{i=1}^{61} \mu_i^{x_i}$, natural selection for translational efficiency w, and genetic drift N_e. Given an expression level ϕ, the probability of observing a set of codons for one amino acid is independent of the probability of observing a set of codons for another amino acid (Section 3.7.1). This independence allows us to calculate the expected frequencies of codons within an amino acid independent of codon compositions of other amino acids. The resulting expected frequency of codon i of amino acid aa_k that has n_k synonymous codons is given by

$$\mathbb{E}[f_i|\phi, aa_k] = \frac{\mu_i e^{-N_e q C \phi t_i}}{\sum_{j \in n_k} \mu_j e^{-N_e q C \phi t_j}}. \tag{3.3}$$

Equation 3 describes how the expected frequency of a given codon changes with gene expression ϕ at its mutation-selection-drift equilibrium. In order to compare our model predictions to observed codon usage frequencies, we looked at the 4674 verified nuclear genes that lack internal stops in *S. cerevisiae* (Gilchrist, 2007; Shah and Gilchrist, 2010b). Since time-average target protein production rates of genes are not available for any organism, we use estimates of protein production rates during log growth as proxies. Empirical estimates of protein production rates ϕ were obtained from (Gilchrist, 2007), which combines mRNA abundance (Beyer *et al.*, 2004) and ribosome occupancy datasets (Arava *et al.*, 2003; MacKay *et al.*, 2004; Shah and Gilchrist, 2010b). The effective population size was set to $N_e = 1.36 \times 10^7$ based on the effective population size of its closely related species *S. paradoxus* (Wagner, 2005). Note that because N_e is scaled by qC in Eqn. 3, any error in our estimate of N_e will only affect our estimates of qC and not the behavior of our predictions.

3.3 Results

3.3.1 Model Behavior.

The general behavior of our model is illustrated in Fig. 3.1, which shows the simple case of one amino acid with two codons. It demonstrates how expected frequencies of the codons change with gene expression with respect to *differences* in the elongation times of the codons $\Delta t_{ij} = t_i - t_j$ as well and their relative mutation rates μ_i/μ_j. As expected, codon usage in genes with low expression is primarily determined by their relative mutation rates, while codon usage in genes with high expression is determined by the differences in their elongation times. When both natural selection for translation efficiency and mutation biases favor the same codon, the lines representing expected frequencies of codons (red lines in Fig. 1) do not cross. However, when the direction of mutation bias is opposite to that of natural selection, the lines representing expected frequencies of codons cross (blue lines in Fig. 3.1).

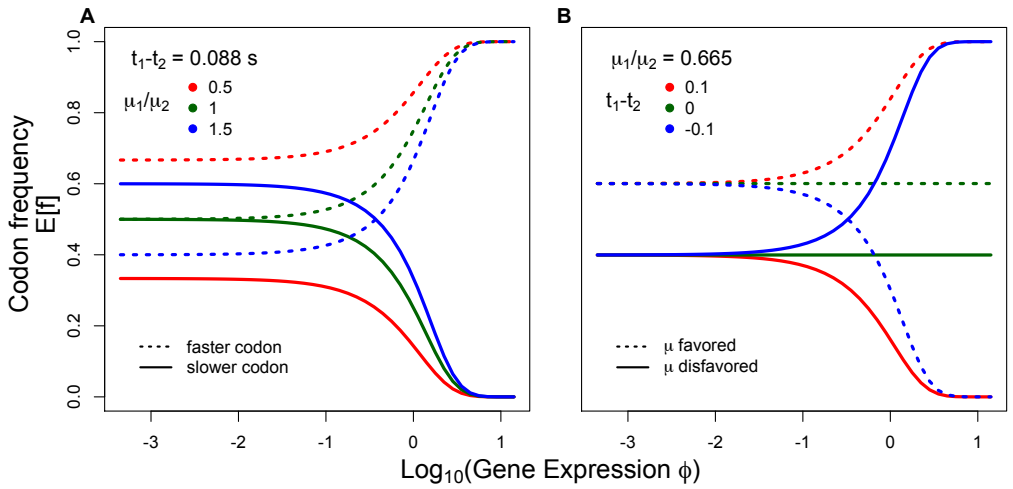


Figure 3.1: Effect of varying relative mutation rates (μ_i/μ_j), elongation times (Δt_{ij}) and protein production rate (ϕ) on the expected codon frequencies ($\mathbb{E}[f]$) in a hypothetical two-codon amino acid.

A Effect of changing μ_i/μ_j on $\mathbb{E}[f]$ with ϕ. Solid lines represent the codon with longer elongation time t_1 and dotted lines represent the codon with shorter elongation time t_2. Mutation bias has a greater effect on $\mathbb{E}[f]$ at low ϕ, while at very high ϕ, the $\mathbb{E}[f]$ of codons converge to the same values irrespective of μ_i/μ_j. **B** Effect of changing $t_i - t_j$ on their expected frequencies $\mathbb{E}[f]$ with respect to ϕ. Solid lines represent the codon with a lower relative mutation rate μ_1 and dotted lines represent the codon with a higher mutation rate μ_2. Differences in elongation times between the two codons $t_1 - t_2$ has little effect on $\mathbb{E}[f]$ at low ϕ. However, at high ϕ, as $t_1 - t_2$ changes, so does the difference in their expected frequencies $\mathbb{E}[f]$.

3.3.2 Model Fit to *S. cerevisiae* Genome.

We calculated maximum likelihood estimates for the composite parameter qC, codon-specific differences in elongation times Δt_{ij}, and relative mutation rates μ_i/μ_j using 4674 genes of the *S. cerevisiae* genome (see Section 3.5, Table 3.1, Table 3.2 for more details).

Table 3.1: Estimates of relative mutation rates μ_i/μ_j

Amino acids	Codons	μ_i/μ_j	Amino acids	Codons	μ_i/μ_j
Ala	μ_{GCC}/μ_{GCA}	0.6541	Pro	μ_{CCC}/μ_{CCA}	0.4460
	μ_{GCG}/μ_{GCA}	0.4016		μ_{CCG}/μ_{CCA}	0.3630
	μ_{GCC}/μ_{GCA}	1.0605		μ_{CCT}/μ_{CCA}	0.8008
Cys	μ_{TGT}/μ_{TGC}	1.6581	Gln	μ_{CAG}/μ_{CAA}	0.5026
Asp	μ_{GAT}/μ_{GAC}	1.9496	Arg	μ_{AGG}/μ_{AGA}	0.5325
Glu	μ_{GAG}/μ_{GAA}	0.4536		μ_{CGA}/μ_{AGA}	0.2012
Phe	μ_{TTT}/μ_{TTC}	1.5262		μ_{CGC}/μ_{AGA}	0.1376
Gly	μ_{GGC}/μ_{GGA}	0.7779		μ_{CGG}/μ_{AGA}	0.1104
	μ_{GGG}/μ_{GGA}	0.5310		μ_{CGT}/μ_{AGA}	0.2946
	μ_{GGT}/μ_{GGA}	1.6471	Ser	μ_{TCC}/μ_{TCA}	0.6861
His	μ_{CAT}/μ_{CAC}	1.8943		μ_{TCG}/μ_{TCA}	0.4736
Ile	μ_{ATC}/μ_{ATA}	0.7647		μ_{TCT}/μ_{TCA}	1.1472
	μ_{ATT}/μ_{ATA}	1.4006		μ_{AGT}/μ_{AGC}	1.4752
Lys	μ_{AAG}/μ_{AAA}	0.6811	Thr	μ_{ACC}/μ_{ACA}	0.6185
Leu	μ_{CTC}/μ_{CTA}	0.4319		μ_{ACG}/μ_{ACA}	0.4740
	μ_{CTG}/μ_{CTA}	0.8441		μ_{ACT}/μ_{ACA}	1.0249
	μ_{CTT}/μ_{CTA}	0.9404	Val	μ_{GTC}/μ_{GTA}	0.7811
	μ_{TTA}/μ_{CTA}	1.9598		μ_{GTG}/μ_{GTA}	0.8533
	μ_{TTG}/μ_{CTA}	1.9253		μ_{GTT}/μ_{GTA}	1.5350
Asn	μ_{AAT}/μ_{AAC}	1.5897	Tyr	μ_{TAT}/μ_{TAC}	1.4217

Table 3.2: Estimates of differences in elongation times Δt (s)

Amino acids	Codons	Δt	Amino acids	Codons	Δt
Ala	$t_{GCC} - t_{GCA}$	-0.1108	Pro	$t_{CCC} - t_{CCA}$	0.1394
	$t_{GCG} - t_{GCA}$	0.0551		$t_{CCG} - t_{CCA}$	0.2514
	$t_{GCC} - t_{GCA}$	-0.1168		$t_{CCT} - t_{CCA}$	0.0396
Cys	$t_{TGT} - t_{TGC}$	-0.0289	Gln	$t_{CAG} - t_{CAA}$	0.1024
Asp	$t_{GAT} - t_{GAC}$	0.0125	Arg	$t_{AGG} - t_{AGA}$	0.1813
Glu	$t_{GAG} - t_{GAA}$	0.0585		$t_{CGA} - t_{AGA}$	0.6795
Phe	$t_{TTT} - t_{TTC}$	0.0419		$t_{CGC} - t_{AGA}$	0.1586
Gly	$t_{GGC} - t_{GGA}$	-0.1452		$t_{CGG} - t_{AGA}$	0.4932
	$t_{GGG} - t_{GGA}$	-0.0593		$t_{CGT} - t_{AGA}$	0.0039
	$t_{GGT} - t_{GGA}$	-0.2126	Ser	$t_{TCC} - t_{TCA}$	-0.0887
His	$t_{CAT} - t_{CAC}$	0.0281		$t_{TCG} - t_{TCA}$	0.0400
Ile	$t_{ATC} - t_{ATA}$	-0.2671		$t_{TCT} - t_{TCA}$	-0.0876
	$t_{ATT} - t_{ATA}$	-0.2588		$t_{AGT} - t_{AGC}$	0.0054
Lys	$t_{AAG} - t_{AAA}$	-0.0443	Thr	$t_{ACC} - t_{ACA}$	-0.0950
Leu	$t_{CTC} - t_{CTA}$	0.1349		$t_{ACG} - t_{ACA}$	0.0600
	$t_{CTG} - t_{CTA}$	0.0733		$t_{ACT} - t_{ACA}$	-0.0902
	$t_{CTT} - t_{CTA}$	0.0674	Val	$t_{GTC} - t_{GTA}$	-0.1736
	$t_{TTA} - t_{CTA}$	-0.0266		$t_{GTG} - t_{GTA}$	-0.0863
	$t_{TTG} - t_{CTA}$	-0.0082		$t_{GTT} - t_{GTA}$	-0.1688
Asn	$t_{AAT} - t_{AAC}$	0.0664	Tyr	$t_{TAT} - t_{TAC}$	0.0683

Although, our model uses $2(k-1)$ parameters for each amino acid with k codons, we show that it is far from being over-parameterized as it uses genome scale datasets (see Section 3.7.2). The fit of our model predictions with observed data is illustrated in Fig. 3.2. Specifically, Fig. 3.2 shows how the observed and predicted codon frequencies change with gene expression ϕ for all the amino acids that use multiple codons. Because the set of synonymous codons for Ser occur in blocks of two and four codons separated by more than a single mutation step, we treat each of the blocks as a separate amino acids, Ser_2 and Ser_4 respectively. The fit of our model can be quantified on a per amino acid basis based on the Pearson correlation ρ_M between mean of binned observed codon frequencies and predicted codon frequencies at mean ϕ value. The ρ_M values ranged from 0.72 to 0.99 with a median value of 0.936.

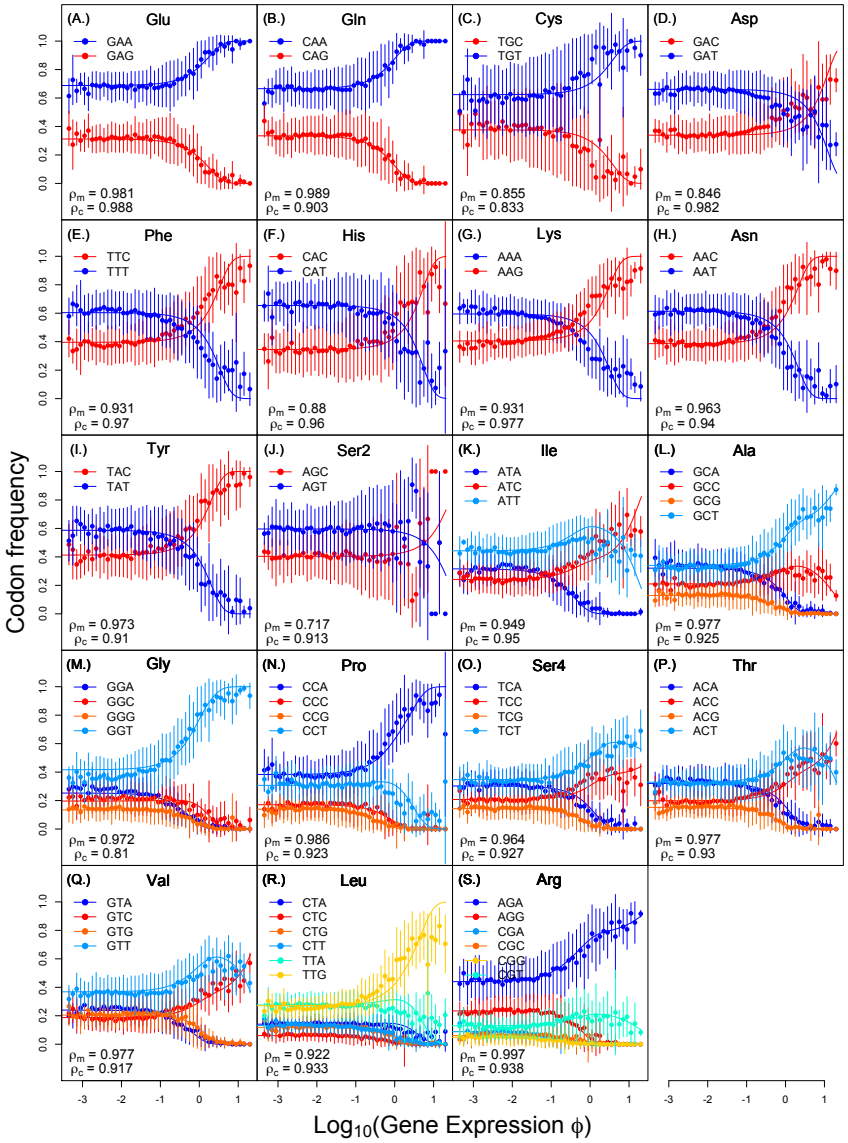


Figure 3.2: Observed and predicted changes in codon frequencies with gene expression, specifically protein production rate ϕ.

Each panel corresponds to a specific amino acid where codons ending in A or T are is shown in shades of blue while codons ending in G or C in shades of red. Solid dots and vertical bars represent mean ± 1 SD of observed codon frequencies within genes with protein production rates defined by the bin. The expected codon frequencies under our model are represented by solid lines. We used $k-1$ codons of an amino acid with k codons in estimating correlation coefficients. ρ_M represents the Pearson correlation between the mean of observed codon frequencies within a bin and predicted codon frequencies at mean ϕ value. ρ_c represents the Pearson correlation between observed codon counts and predicted codon counts of all genes at their specific ϕ value.

Although many indices of adaptation have been proposed to estimate the degree of codon bias within a gene, there exists no method or index that makes predictions on codon counts of individual genes itself. For instance, if a particular gene has a protein production rate ϕ, what should the distribution of its codons be given its amino acid sequence? In order to directly address this question we used our estimates of Δt_{ij} and μ_i/μ_j (Table 3.1, Table 3.2) to evaluate on a per-gene basis the expected codon frequencies for each amino acid using Eqn. 3. We find that the correlation between observed and predicted codon counts is $\rho_c = 0.959$ (Fig. 3.3), explaining $\sim 92\%$ of observed variation in codon counts. Even at the level of individual amino acids, the correlation coefficients ρ_c ranged from $0.81 - 0.99$. All but two amino acids had $\rho_c > 0.9$, indicating that the high correlation was consistent across all amino acids. In summary, we find that our model does an excellent job of predicting how the observed codon frequencies in *S. cerevisiae* change with gene expression ϕ.

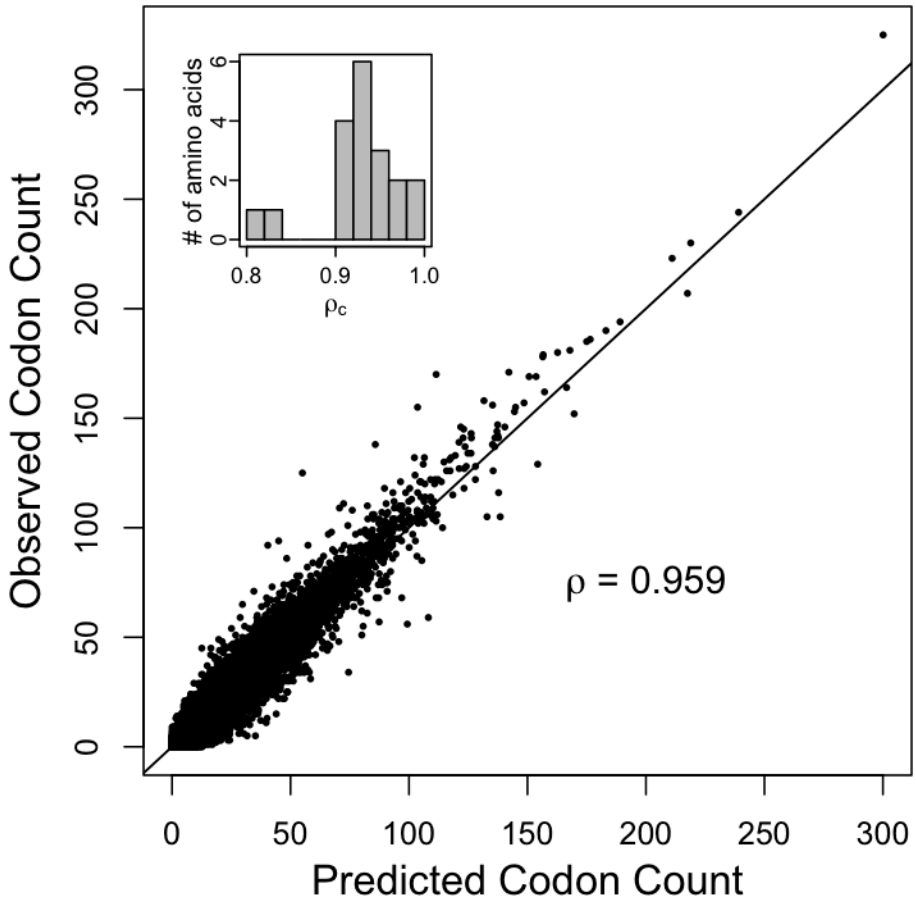


Figure 3.3: Correlation between observed codon counts and predicted codon counts of individual genes.

We used codon counts of $k-1$ codons of an amino acid with k codons. Ignoring Met and Trp (one codon amino acids) and splitting Ser into two blocks of four and two codons, there are 19 unique amino acid sets. Hence the number of data points used are $4674 \times (59 - 19) = 186,960$. We find a very high correlation ($\rho = 0.959$, p-value $< 10^{-15}$) between our model predictions and observed counts. Inset shows the distribution of correlation coefficients at the level of individual amino acids, indicating that our high correlation is not biased by specific amino acids and that we have a high correlation across all amino acids.

One key insight from this work is that in *S. cerevisiae* for amino acids with more than two codons, the frequencies of preferred codons with similar elongation times Δt_{ij} can change in a non-monotonic manner with gene expression ϕ. For instance, in the case of Thr, the frequency of codon ACT increases from low to moderate levels of gene expression $\log(\phi)$ but decreases at high gene expression and is replaced by codon ACC. This non-monotonic behavior is the result of complex interplay between mutation biases and translation selection. Specifically, although both the codons ACC and ACT have shorter elongation times than their other coding synonyms ACG and ACA, codon ACC has the shortest elongation time. However, unlike codon ACC, ACT is favored by mutation bias, so its frequency initially increases with gene expression. We call this phenomenon 'mutational inertia', whereby, the frequency of a suboptimal codon transiently increases with gene expression due to mutation bias. This non-monotonic behavior runs counter to traditional explanations where the frequency of an optimal codon is expected to monotonically increase and that of a suboptimal codon to monotonically decrease with gene expression (Sharp and Li, 1986; Duret and Mouchiroud, 1999). We observed these effects of mutational inertia in most of the amino acids with more than two codons. Although non-monotonic changes in codon frequencies with gene expression have been previously documented (Bulmer, 1988), the mechanisms responsible for this behavior have not been put forth. We believe this interesting and complex interplay between mutation biases and selection for efficient translation has been obscured due to an overemphasis on indices in studies of codon usage bias. Our study illustrates the advantages of model-based approach used here over heuristic approaches. In addition and as indicated by the crossing of lines representing codon frequencies, 7 out of 10 amino acids with two codons in Fig. 3.2 (D-J), show mutation biases in a direction opposite to that of natural selection. In other words, codons with high frequencies in low expression genes are not the same as the ones preferred in high expression genes. Along with explaining these previously described patterns (Sharp and Devine, 1989; Musto *et al.*, 2003; Peixoto *et al.*, 2004), we quantity the changes in codon frequencies with gene expression.

In addition to describing the genome scale patterns of codon usage, our model also allows for estimation of relative mutation rates μ_i/μ_j and differences in elongation times of these codons Δt_{ij} on a per amino acid basis directly from the genome sequence and expression datasets. Interestingly, we find that estimates of relative mutation rates sometimes differed between amino acids. For instance, in the case of two-codon amino acids (Lys, Gln, and Glu) the NNA codons were always favored over NNG codons. However, the relative mutation rate μ_{NNG}/μ_{NNA} ranged from 0.45-0.68 with a mean of 0.546. These small but significant differences (t test, $p < 10^{-9}$ for every pair of amino acids) in the estimation of relative mutation rate may be due, in part, to the fact that our model does not allow for non-synonymous substitutions, some of which may behave in a nearly neutral manner, especially in genes with low ϕ values.

We also compared our estimates of Δt_{ij} with estimates based on tRNA gene copy numbers as proxy for tRNA abundances and wobble penalties (see Methods). We find that these independently obtained estimates of Δt_{ij} are highly correlated ($\rho = 0.801$) (Fig. 3.4).

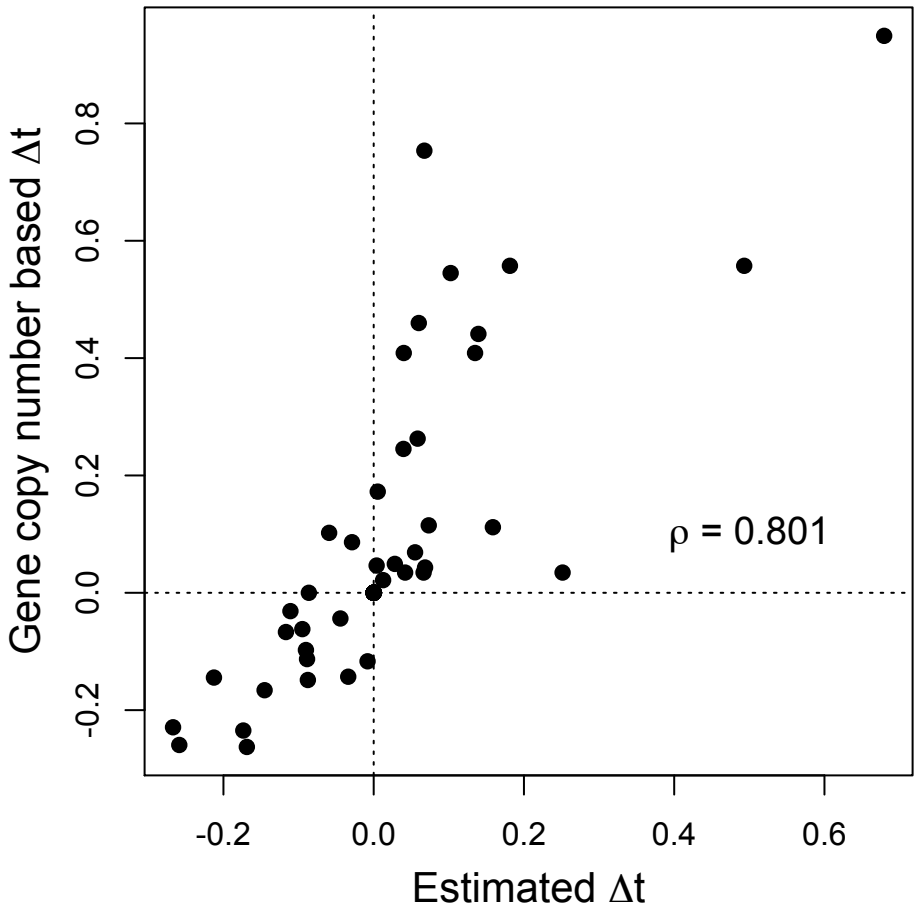


Figure 3.4: Correlation between our model based estimates of Δt_{ij}s with Δt_{ij}s estimated using tRNA gene copy numbers.

We find a strong correlation ($\rho = 0.801$, p-value $< 10^{-9}$) between our model estimates and estimates of Δt_{ij} based on tRNA gene copy numbers indicating that our estimates can be related to other biological estimates such as tRNA abundances directly.

3.3.3 Model Fit vs. Model Predictions.

In order to demonstrate the predictive value of our model, we randomly partitioned the *S. cerevisiae* genome into two sets of 2337 genes each with no signifiant bias in their distribution of gene expression levels ϕ (t test, $p > 0.4$). Parameters estimated

using half the genome were found to be highly correlated with our previous estimates based on the entire genome $\rho > 0.99$ for both Δt_{ij} and μ_i/μ_j (Fig. 3.5).

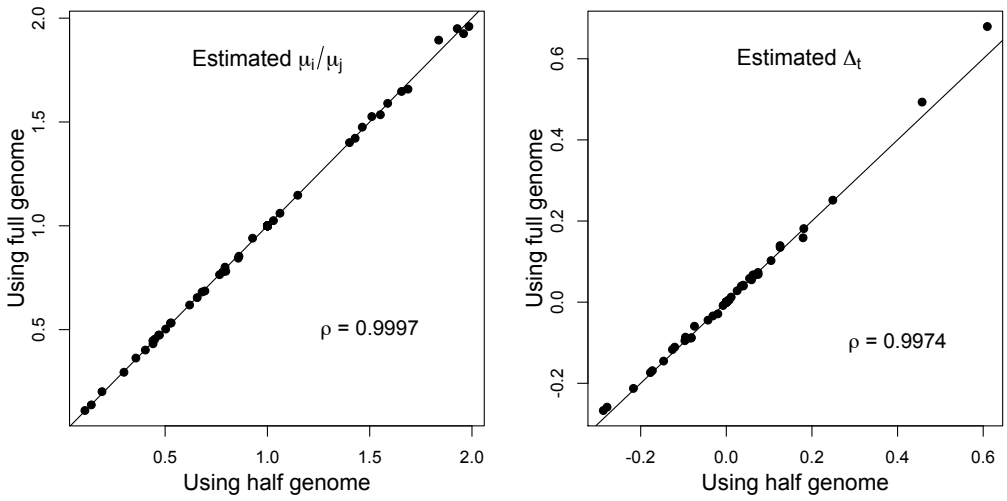


Figure 3.5: Correlation between estimates of Δts andμ_i/μ_j using a random subset of 2337 genes (half the genome) and using the entire genome.
We find a strong correlation ($\rho > 0.99$, p-value $< 10^{-15}$) for both Δt and μ_i/μ_j.

We then used the parameters estimated using the first set of genes to predict gene-specific codon counts in the second set of genes. The correlation coefficient between observed and predicted codon counts at the level of individual genes was 0.96 (Figs. 3.6 and 3.7).

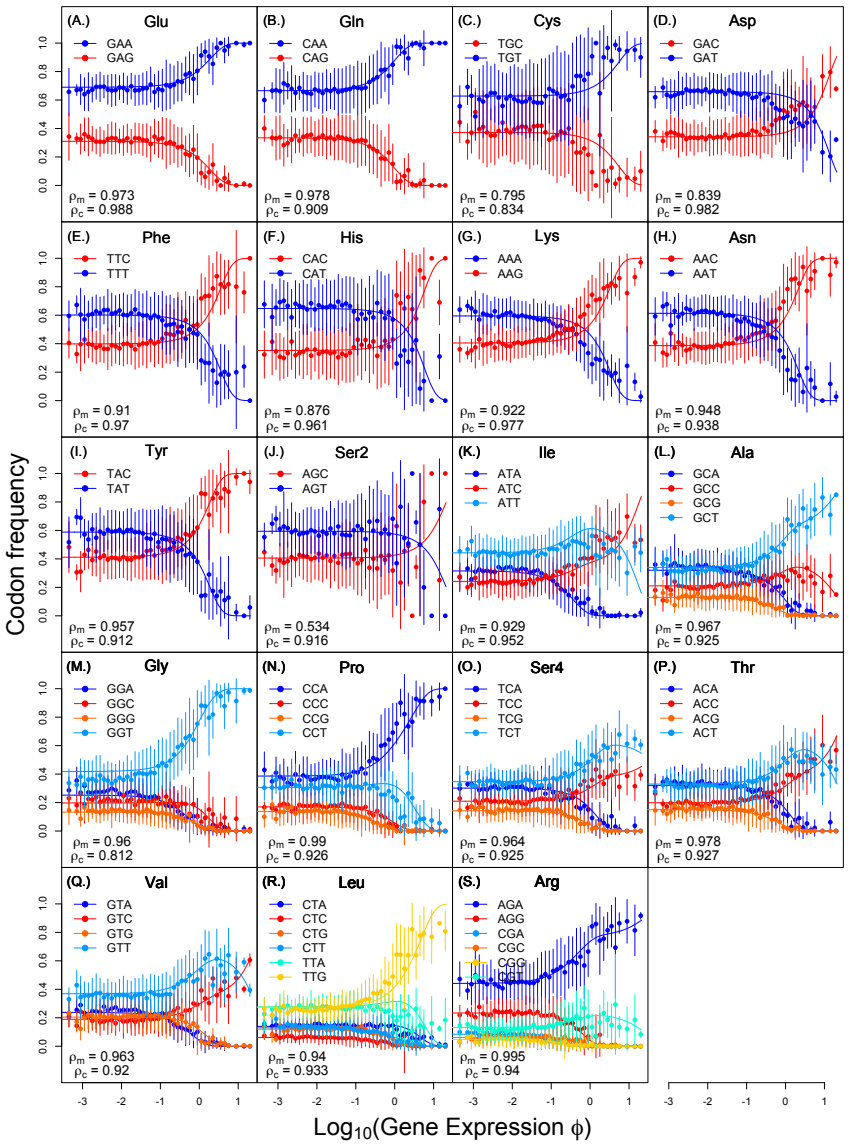


Figure 3.6: Observed and predicted changes in codon frequencies with gene expression for the second half of the genome using parameters Δt and μ_i/μ_j estimated using the first half.

Each panel corresponds to a specific amino acid where codons ending in A/T are is shown in shades of blue while codons ending in G/C in shades of red. Solid dots and vertical bars represent mean ± 1 SD of observed codon frequencies within genes with protein production rates defined by the bin. The expected codon frequencies under our model are represented by solid lines. ρ_M represents the correlation between the mean of observed codon frequencies in a bin and predicted codon frequencies at mean ϕ value. ρ_c represent the correlation between observed codon counts and predicted codon counts of all genes at their specific ϕ value.

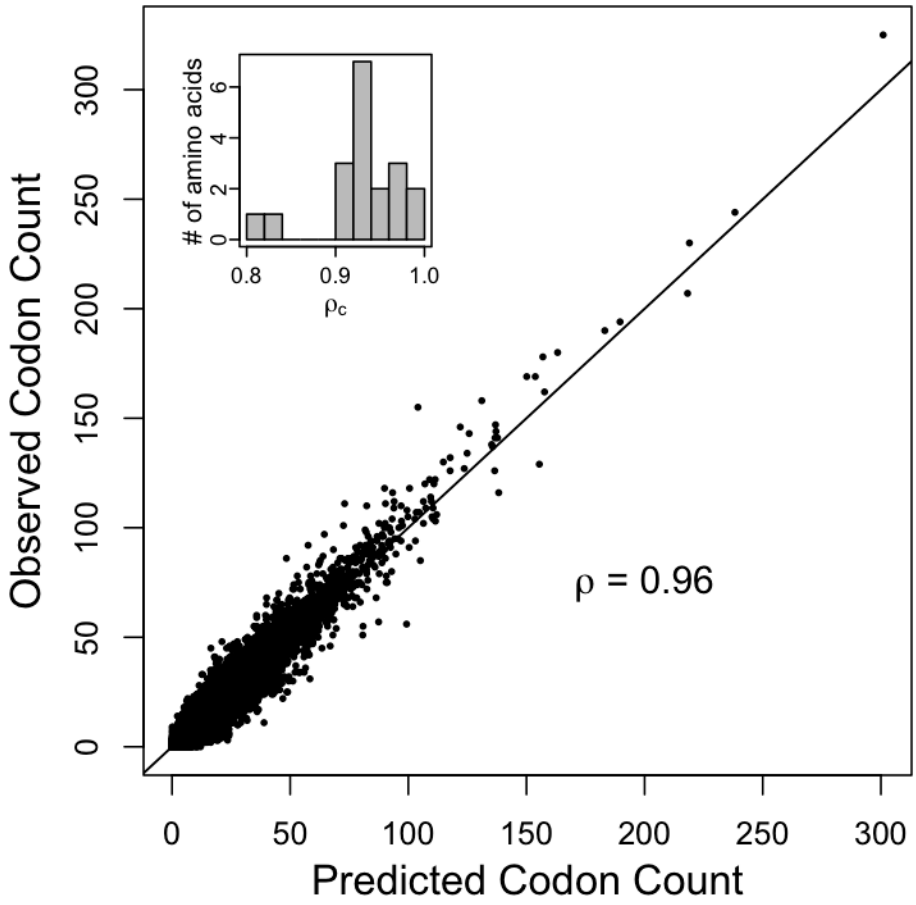


Figure 3.7: Correlation between observed codon counts and predicted codon counts of individual genes in second half of the genome using parameters Δt and μ_i/μ_j estimated using the first half.
We find a very high correlation ($\rho = 0.96$, p-value $< 10^{-15}$) between our model predictions and observed counts. Inset shows the distribution of correlation coefficients at the level of individual amino acids, indicating that our high correlation is not biased by specific amino acids and that we have a high correlation across all amino acids.

Since for most organisms we do not have ribosome occupancy datasets to estimate protein production rates, we estimated Δt_{ij} and μ_i/μ_j using mRNA abundances

(Beyer *et al.*, 2004; Gilchrist, 2007) as proxies for protein production rates ϕ. We found a very high correlation between parameters estimated using mRNA abundances and protein production rates ($\rho > 0.97$, Figs. 3.8 and 3.9). Because our model is based on mechanistic principles of protein translation, these parameters can be directly related to specific biological processes underlying protein translation. Our work demonstrates that, in principle, these parameters can be estimated directly from genomic and expression datasets, as shown above. Estimation of these parameters can thus be easily extended to any sequenced organisms for which genome scale expression datasets exist.

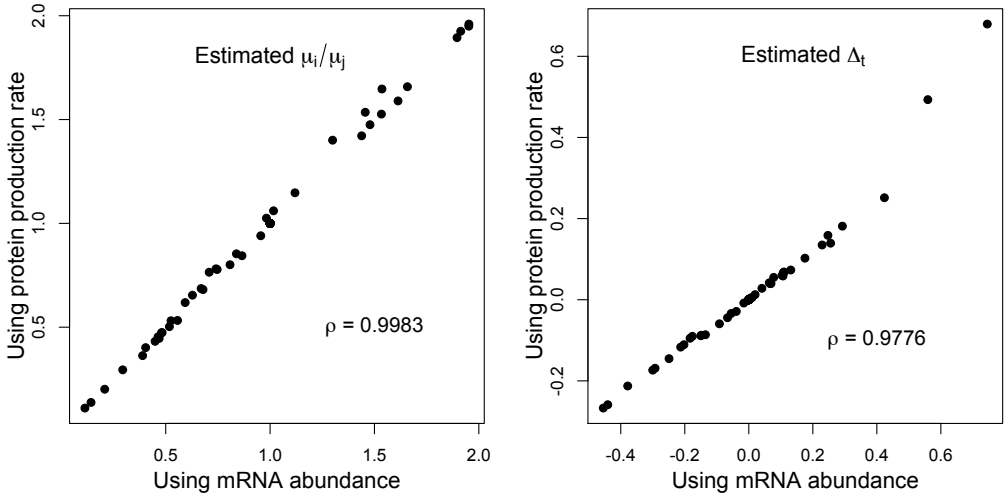


Figure 3.8: Correlation between estimates of Δts andμ_i/μ_j using protein production rates ϕ for each gene and using mRNA abundances.
We find a strong correlation ($\rho > 0.97$, p-value $< 10^{-15}$) for both Δt and μ_i/μ_j.

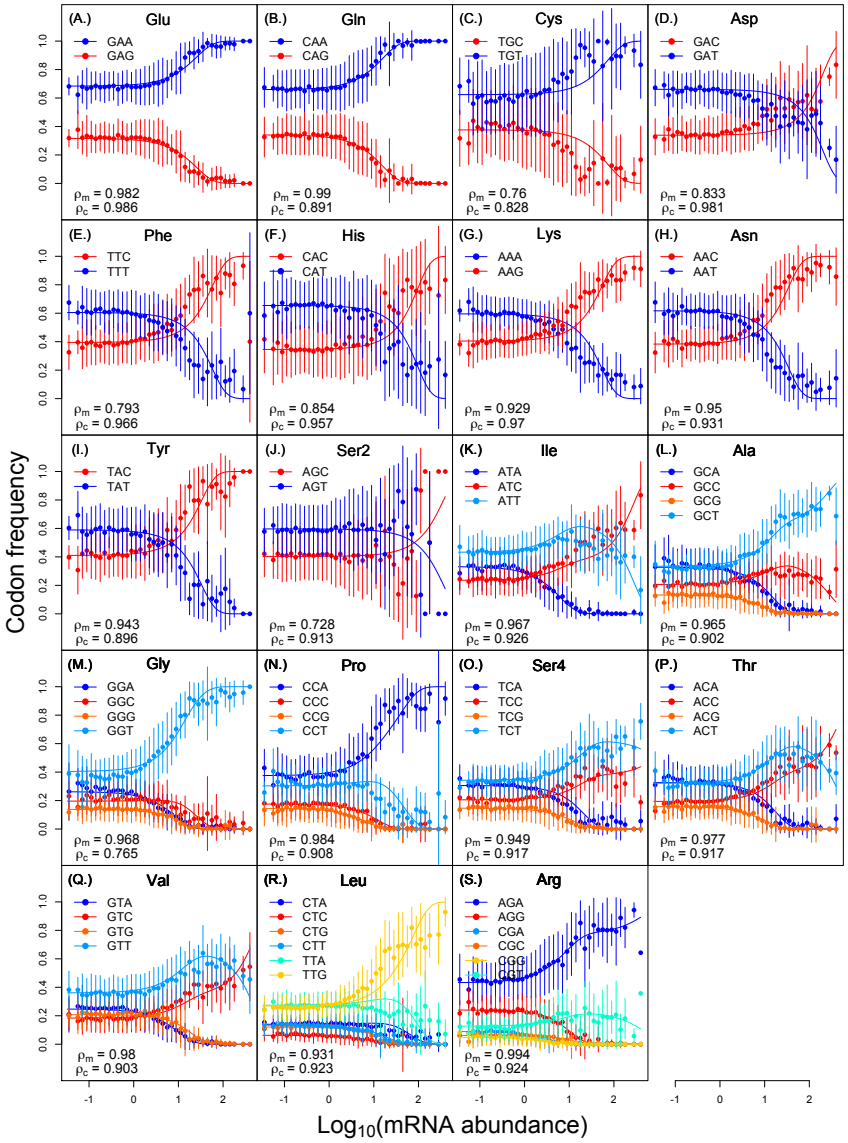


Figure 3.9: Observed and predicted changes in codon frequencies with gene expression, specifically mRNA abundances.

Each panel corresponds to a specific amino acid where codons ending in A/T are is shown in shades of blue while codons ending in G/C in shades of red. Solid dots and vertical bars represent mean ±1 SD of observed codon frequencies within genes with mRNA abundances defined by the bin. The expected codon frequencies under our model are represented by solid lines. ρ_M represents the correlation between the mean of observed codon frequencies in a bin and predicted codon frequencies at mean mRNA abundance of the bin. ρ_c represent the correlation between observed codon counts and predicted codon counts of all genes at their specific ϕ value.

3.4 Discussion

3.4.1 Broader Interpretation of Δt_{ij}.

The high correlation between estimates of Δt_{ij} from independent sources of genomic information (Fig. 3.4), suggests that our interpretation of the term Δt_{ij} is consistent with selection for translation efficiency as a major force in shaping patterns of codon usage. However, from a purely mathematical standpoint, the parameter Δt_{ij} is akin to the additive fitness component used in (Sella and Hirsh, 2005), scaled by ϕ. Thus its value can broadly be interpreted as an expression level dependent selective coefficient associated with the specific codon pair. In future, this broader interpretation should allow us to compare our genome-based estimates of Δt_{ij} with values expected under alternate hypotheses of the factors responsible for shaping codon usage patterns. For example, in the case of Cys, an interpretation of Δt_{ij} is difficult to justify based on a naive model of estimating elongation times from tRNA abundances. In *S. cerevisiae*, Cys is coded by a single tRNA where the non-canonical codon TGT is recognized by wobble and assumed to be elongated at a slower rate than its synonym TGC (Gromadski and Rodnina, 2004; Shah and Gilchrist, 2010b). Thus, our estimates of $t_{\mathrm{TGT}} - t_{\mathrm{TGC}} < 0$ cannot be explained on the basis of elongation times alone as the sign of $\Delta t_{TGT,TGC}$ is opposite to that expected based on tRNA abundances and wobble. A variety of factors could potentially explain this discrepancy. Firstly, due to its unique ability to form disulphide linkages, Cys might be under a stronger selection to minimize missense errors than other amino acids. The fact that a codon with a slower elongation rate might be better at minimizing missense errors has also been predicted in a large number of other microorganisms (Shah and Gilchrist, 2010b). Secondly, as noted by (Bennetzen and Hall, 1982), codons with side-by-side GC nucleotides may be selected against due to the high binding energies between codon-anticodon pairs. Despite the fact that Δt_{ij} can potentially be interpreted many ways, the high correlation between our predicted

Δt_{ij} and estimates of Δt_{ij} based simply on tRNA gene copy numbers and wobble parameters (Fig. 3.4) indicates a mechanistic link between our estimates of Δt and differences in elongation times of codons.

In summary, our work shows that genome scale patterns of codon usage can be largely explained by the effects of genetic drift, mutational biases, and natural selection for efficient usage of ribosome, i.e. translational efficiency. Although a variety of indices have been proposed to estimate the degree of adaptation of a gene based on its codon usage bias, ours method makes predictions in the opposite direction as well, i.e., predicting codon counts of a gene given its expression level. Our model of translation efficiency also allows us to estimate codon-specific elongation times (selection coefficients) as well as relative mutation rates. In addition, we make quantitative predictions on how individual codon frequencies should change with gene expression in yeast. Although, selection for translational efficiency appears to be sufficient to explain most of the genome-scale patterns of codon usage this does not preclude the effects of other selective forces on the evolution of CUB. For instance, selection for translation accuracy (minimizing translation missense errors) has long been argued to be a dominant force in driving the evolution of CUB (Akashi, 1994; Drummond *et al.*, 2005; Drummond and Wilke, 2008). However, current data suggests that only $\sim 10-50\%$ of missense errors disrupt protein function (Markiewicz *et al.*, 1994; Guo *et al.*, 2004), and therefore cannot explain the high frequencies $\sim$100% of mutationally disfavored codons in Phe, Asn, and Tyr amino acids (Fig. 3.2). Moreover, the assumptions underlying Akashis test (Akashi, 1994) used to support the translation accuracy hypothesis are not always justified (Shah and Gilchrist, 2010b). Nevertheless, selection for translation accuracy can explain codon usage at functionally and/or structurally critical sites of a protein (Drummond and Wilke, 2008). Because codons that minimize missense errors may not necessarily be the ones that minimize elongation times (Shah and Gilchrist, 2010b), our model is likely insufficient to explain the codon usage at these sites. Similarly, adaptation against nonsense errors has been documented in *S. cerevisiae*

(Gilchrist and Wagner, 2006; Gilchrist *et al.*, 2009) and other organisms (Qin *et al.*, 2004). In addition, factors indirectly related to protein translation, such as mRNA secondary structures at the 5′ region of a gene, have been shown to be under selection for translation initiation and hence can effect the frequency of codon usage at these sites (Kudla *et al.*, 2009; Tuller *et al.*, 2010).

Clearly, although a number of selective mechanisms have been proposed to explain and likely contribute to specific patterns of codon usage, the combined effects of these forces in shaping genomic patterns of codon usage are not well understood (Drummond and Wilke, 2009; Plotkin and Kudla, 2011). In order to decipher the relative importance of these forces on the evolution of CUB, mechanistic models that explicitly take into account tRNA competition and intra-ribosomal dynamics (Shah and Gilchrist, 2010b) as well as effects of amino acid substitutions on protein structure and function (Guo *et al.*, 2004) need to be developed. Our model demonstrates the strength of such an approach and provides a natural framework for expansion to include other selective forces as well. More generally, this approach will allow us to quantitatively estimate parameters underlying fundamental biological processes such as protein translation and improve our understanding of how evolutionary forces shape genomic patterns and processes.

3.5 Methods

3.5.1 Estimation of Δt_{ij} and μ_i/μ_j from observed data

In the case of an amino acid with k codons, the change in codon frequencies across the entire range of gene expression can be determined by $2(k-1)$ parameters for codon-specific mutation rates and elongation times. For instance, in the case of amino acids with two codons, the frequency of any one codon depends only on the *difference* in

the elongation times of the two codons and the ratio of their mutation rates.

$$\begin{aligned}\mathbb{E}[x_1|\phi] &= \frac{n\mu_1 e^{-N_e qC\phi t_1}}{\mu_1 e^{-N_e qC\phi t_1} + \mu_2 e^{-N_e qC\phi t_2}} \\ &= \frac{1}{1 + \frac{\mu_2}{\mu_1} e^{-N_e qC\phi (t_2 - t_1)}}\end{aligned} \tag{3.4}$$

Codon usage in genes with low expression ϕ is thought to be determined primarily by mutation biases, i.e., $N_e qC\phi \approx 0$. Since absolute mutation rates to each codon cannot be estimated directly as it is only their ratios that affect codon usage, we estimated μ_i/μ_j by setting the mutation rate of an arbitrarily chosen codon to 1. Codon counts in low expression genes can then be assumed to follow a multinomial distribution with parameters determined by their mutation rates. Thus, in the case of an amino acid with two codons whose codon counts are x_1 and x_2, the maximum likelihood estimate of relative mutation rate is approximately,

$$\frac{\mu_2}{\mu_1} \approx \frac{x_2}{x_1} \tag{3.5}$$

Similarly, elongation times of codons affect codon usage only as their differences $(t_1 - t_2)$. Thus, during parameter estimation of elongation times, we set the elongation time of an arbitrarily chosen codon within each amino acid to 1 and estimated the differences in elongation times of other codons with respect to that codon. We used the NEWUOA optimization algorithm (Powell, 2006) employed in R to estimate Δt_{ij} and μ_i/μ_j for an amino acid with k codons and qC by maximizing the following likelihood function (see Section 3.7.1 for additional details).

$$\text{Lik}(\vec{t}, \vec{\mu}|\phi, \vec{x}) = P(\vec{x}|\phi) = \prod_{i=1}^{k} \left(\frac{\mu_i e^{-N_e qC\phi t_i}}{\sum_{j=1}^{k} \mu_j e^{-N_e qC\phi t_j}} \right)^{x_i} \tag{3.6}$$

In addition, we estimated the maximum likelihood value of $\widehat{qC} = 9.12 \times 10^{-7}$.

3.5.2 Estimation of Δt_{ij} from tRNA gene copy numbers

In order to compare our estimates of Δt_{ij} with an independent source of genomic information, we estimated Δt_{ij} using tRNA gene copy numbers and wobble effects. Following (Dong *et al.*, 1996; Kanaya *et al.*, 1999), we use tRNA gene copy numbers in yeast obtained from GtRNAdb (Chan and Lowe, 2009) as proxies for tRNA abundances. We assume that the expected waiting time at a codon t_i is inversely proportional to its cognate tRNA abundances based on an exponential waiting process.

$$[\text{tRNA}_i] \propto \text{Gene copy number of tRNA}_i \tag{3.7}$$

$$t_i = \frac{a}{[\text{tRNA}_i] \times wob} \tag{3.8}$$

where wob is the wobble penalty due to codon-anticodon mismatch and a is a scaling constant. When a codon is recognized by its canonical tRNA, we set $wob = 1$. Based on (Curran and Yarus, 1989; Lim and Curran, 2001), we assume that a purine-purine or pyrimidine-pyrimidine wobble penalty to be 39% and purine-pyrimidine wobble penalty to be 36%. We set the scaling constant a such that the harmonic mean of elongation rates of all codons is 10 aa/sec (Gilchrist and Wagner, 2006; Gilchrist, 2007). However, note that changing the scaling constant would have no effect on the correlation between our model based and gene copy number based estimates of Δt_{ij}.

3.6 Acknowledgements

We thank J. Plotkin, B. O'Meara, F. Ubeda de Torres, I. Juric for their comments on the manuscript. Support for this project was provided by Department of Ecology and Evolutionary Biology at University of Tennessee, Knoxville, the National Institute for Mathematical and Biological Synthesis (NIMBioS), and the TN Science Alliance. P.S. was additionally funded by a NIMBioS Graduate Research Assistantship.

3.7 Supporting Information

3.7.1 Analytical solutions of the model

One amino acid with two codons

Consider a gene sequence of length n composed of a single two-codon amino acid, whose average elongation times are t_1 and t_2. Let x_1 and $x_2 = n - x_1$ be the respective codon counts. The expected cost of ribosome usage during protein production is then given as

$$\eta(\vec{x}) = C \sum_{i=1}^{2} x_i t_i \tag{3.9}$$

$$= C(x_1 t_1 + x_2 t_2) \tag{3.10}$$

where C is the cost of ribosome usage in ATP/sec. We assume an exponential fitness function w described as

$$w(\vec{x}|\phi) = e^{-q\phi\eta(\vec{x})} = e^{-q\phi C(x_1 t_1 + x_2 t_2)} \tag{3.11}$$

where ϕ is the protein production rate, a measure of gene expression and q is the scaling constant determining the relationship between cost of ATP usage to organismal fitness w.

Following (Kimura, 1964; Gavrilets, 2004; Berg *et al.*, 2004; Sella and Hirsh, 2005), the probability of observing an allele across the entire genotype space at equilibrium is given by

$$P(\vec{x}|\phi) = \frac{w(\vec{x}|\phi)^{N_e}}{\sum_{y \in S_c} w(\vec{y}|\phi)^{N_e}} \tag{3.12}$$

where N_e is the effective population size and S_c is the entire synonymous codon genotype space, which has 2^n alleles in this simple case. Since the cost of protein

production is independent of codon order within a gene, multiple synonymous alleles could give rise to the same cost η. In the 2 codon case, the number of alleles with the same cost is represented by a binomial coefficient and for amino acids with more than two codons, the combinations will be represented by a multinomial coefficient.

$$P(\vec{x}|\phi) = \frac{\binom{n}{x_1} e^{-N_e q \phi C(x_1 t_1 + x_2 t_2)}}{\sum_{y_1=0}^{n} \binom{n}{y_1} e^{-N_e q \phi C(y_1 t_1 + y_2 t_2)}} \tag{3.13}$$

Let μ_1 and μ_2 represent the rate of mutations *to* the two codons as described by (Sella and Hirsh, 2005). For instance, $\mu_1 = \mu_{21}$ indicates the rate at which codon 2 is mutated to codon 1.

Taking mutational biases into account, the probability of observing a given allele is given as

$$P(\vec{x}|\phi) \propto w(\vec{x}|\phi)^{N_e} \prod_{i=1}^{2} \mu_i^{x_i} \tag{3.14}$$

$$P(\vec{x}|\phi) = \frac{\binom{n}{x_1} e^{-N_e q C \phi(x_1 t_1 + x_2 t_2)} \prod_{i=1}^{2} \mu_i^{x_i}}{\sum_{y_1=0}^{n} \binom{n}{y_1} e^{-N_e q C \phi(y_1 t_1 + y_2 t_2)} \prod_{i=1}^{2} \mu_i^{y_i}} \tag{3.15}$$

where $\vec{x} = \{x_1, x_2\}$.

Given the protein production rate ϕ (gene expression) of a gene and the elongation times t of codons, the expected count of each codon is given as

$$\mathbb{E}[x_1|\phi] = \sum_{x_1=0}^{n} x_1 P(\vec{x}|\phi) \tag{3.16}$$

$$= \sum_{x_1=0}^{n} x_1 \frac{\binom{n}{x_1} e^{-N_e q C \phi(x_1 t_1 + x_2 t_2)} \prod_{i=1}^{2} \mu_i^{x_i}}{\sum_{y_1=0}^{n} \binom{n}{y_1} e^{-N_e q C \phi(y_1 t_1 + y_2 t_2)} \prod_{i=1}^{2} \mu_i^{y_i}} \tag{3.17}$$

$$= \frac{n \mu_1 e^{-N_e q C \phi t_1}}{\mu_1 e^{-N_e q C \phi t_1} + \mu_2 e^{-N_e q C \phi t_2}} \tag{3.18}$$

and by symmetry

$$\mathbb{E}[x_2|\phi] = \frac{n\mu_2 e^{-N_e qC\phi t_1}}{\mu_1 e^{-N_e qC\phi t_1} + \mu_2 e^{-N_e qC\phi t_2}} \tag{3.19}$$

$$= n - \mathbb{E}[x_1|\phi] \tag{3.20}$$

One amino acid with k codons

Using the methods described above it can be showed that for any amino acid with k codons, the expected count of the i^{th} codon is given as

$$\mathbb{E}[x_i|\phi] = \frac{n\mu_i e^{-N_e qC\phi t_i}}{\sum_{j=1}^{k} \mu_j e^{-N_e qC\phi t_j}} \tag{3.21}$$

Thus, the expected frequencies of each codon $f_i = x_i/n$ is given as

$$\mathbb{E}[f_i|\phi] = \frac{\mu_i e^{-N_e qC\phi t_i}}{\sum_{j=1}^{k} \mu_j e^{-N_e qC\phi t_j}} \tag{3.22}$$

Variance around the expected value $\mathbb{E}x_i|\phi$ can also be calculated as

$$\mathrm{Var}[x_i|\phi] = \sum_{x_i=0}^{n} (x_i - \mathbb{E}x_i|\phi)^2 P(\{x_1, x_2, \cdots, x_k\}) \tag{3.23}$$

$$= \frac{n\left(\prod_{j=1}^{k} \mu_j\right) e^{N_e qC\phi \sum_{j=1}^{k} t_j}}{\left(\sum_{j=1}^{k} \mu_j e^{N_e qC\phi t_j}\right)^2} \tag{3.24}$$

Multiple amino acids with varying number of codons

In the case of real genes, which are comprised of multiple amino acids each with a varying number of codons, the expected counts and frequencies of codons can be estimated from the marginal distributions of each amino acid. For instance, consider the simple case of two amino acids with two codons each. The ribosomal overhead

cost of protein production is given as

$$\eta(\vec{x}) = C(x_{11}t_{11} + x_{12}t_{12} + x_{21}t_{21} + x_{22}t_{22}) \tag{3.25}$$

where x_{ij} is the number of codons of type j of amino acid i in the gene. Let $n_1 = x_{11} + x_{12}$ and $n_2 = x_{21} + x_{22}$ be the counts of the two amino acids in the gene. As earlier, the probability of observing an allele can be written as

$$P(\vec{x}|\phi) = \frac{\binom{n_1}{x_{11}}\binom{n_2}{x_{21}}\prod_{j=1}^{2}\mu_{1j}^{x_{1j}}\prod_{j=1}^{2}\mu_{2j}^{x_{2j}}e^{-N_e(x_{11}qC\phi t_{11}+x_{12}qC\phi t_{12}+x_{21}qC\phi t_{21}+x_{22}qC\phi t_{22})}}{\sum_{y_{11}=0}^{n_1}\sum_{y_{21}=0}^{n_2}\binom{n_1}{y_{11}}\binom{n_2}{y_{21}}\prod_{j=1}^{2}\mu_{1j}^{y_{1j}}\prod_{j=1}^{2}\mu_{2j}^{y_{2j}}e^{-N_e(y_{11}qC\phi t_{11}+y_{12}qC\phi t_{12}+y_{21}qC\phi t_{21}+y_{22}qC\phi t_{22})}} \tag{3.26}$$

$$= \frac{\binom{n_1}{x_{11}}\prod_{j=1}^{2}\mu_{1j}^{x_{1j}}e^{-N_e(x_{11}qC\phi t_{11}+x_{12}qC\phi t_{12})}}{\sum_{y_{11}=0}^{n_1}\binom{n_1}{y_{11}}\prod_{j=1}^{2}\mu_{1j}^{x_{1j}}e^{-N_e(x_{11}qC\phi t_{11}+x_{12}qC\phi t_{12})}} \times \frac{\binom{n_2}{x_{21}}\prod_{j=1}^{2}\mu_{2j}^{x_{2j}}e^{-N_e(x_{21}qC\phi t_{21}+x_{22}qC\phi t_{22})}}{\sum_{y_{21}=0}^{n_2}\binom{n_2}{y_{21}}\prod_{j=1}^{2}\mu_{2j}^{x_{2j}}e^{-N_e(x_{21}qC\phi t_{21}+x_{22}qC\phi t_{22})}} \tag{3.27}$$

$$P(\{\vec{x}_1, \vec{x}_2\}) = P(\vec{x}_1|aa_1)P(\vec{x}_2|aa_2) \tag{3.28}$$

The marginal distribution of genotype space of a singe amino acid is given as

$$\sum_{x_{21}=0}^{n_2} P(\vec{x}_2|aa_2) = 1 \tag{3.29}$$

$$P(\vec{x}_1|aa_1) = \sum_{x_{21}=0}^{n_2} P(\{\vec{x}_1, \vec{x}_2\}) \tag{3.30}$$

Thus, the expected number of codons of a specific amino acid based on the marginal distribution of that amino acid can be calculated as

$$\mathbb{E}[x_{11}|\phi] = \sum_{x_{11}=0}^{n_1} x_{11} \sum_{x_{21}=0}^{n_2} P(\{\vec{x}_1, \vec{x}_2\}) \tag{3.31}$$

$$= \sum_{x_{11}=0}^{n_1} x_{11} P(\vec{x}_1|aa_1) \sum_{x_{21}=0}^{n_2} P(\vec{x}_2|aa_2) \tag{3.32}$$

$$= \sum_{x_{11}=0}^{n_1} x_{11} P(\vec{x}_1|aa_1) \tag{3.33}$$

$$= \frac{n_1 \mu_{11} e^{-N_e q C \phi t_{11}}}{\mu_{11} e^{-N_e q C \phi t_{11}} + \mu_{12} e^{-N_e q C \phi t_{12}}} \tag{3.34}$$

The above Eqn. (27) is equivalent to Eqn. (11) which considers a gene sequence with only one amino acid and two codons.

3.7.2 An argument against model over-parametrization

Although, it may seem that the excellent fit between the observed and predicted values may be due to over-fitting the data with a large numbers of parameters, this is not the case. For instance, in the case of an amino acid with k codons, there are $k-1$ independent codon frequencies. Since the change in codon frequencies with gene expression can be thought of as a non-linear regression, each codon should have a slope and an intercept. Thus there are $2(k-1)$ independent parameters for an amino acid with k codons. The relative mutation rates provide the estimates for intercepts, while differences in elongation times provide the estimates for their respective slopes. The beauty of our approach lies in the fact that our simple model, appropriately parameterized leads to a correlation coefficient of 0.96.

Chapter 4

Is thermosensing property of RNA thermometers unique?

This chapter is a lightly revised version of a paper by the same name published in PLoS ONE and co-authored with Michael A. Gilchrist.

Shah and Gilchrist. Is thermosensing property of RNA thermometers unique?. PLoS ONE (2010) vol. 5 (7) pp. e11308

Abstract

A large number of studies have been dedicated to identify the structural and sequence based features of RNA thermometers, mRNAs that regulate their translation initiation rate with temperature. It has been shown that the melting of the ribosome-binding site (RBS) plays a prominent role in this thermosensing process. However, little is known as to how widespread this melting phenomenon is as earlier studies on the subject have worked with a small sample of known RNA thermometers. We have developed a novel method of studying the melting of RNAs with temperature by computationally sampling the distribution of the RNA structures at various temperatures using the RNA folding software Vienna. In this study, we compared the thermosensing property of 100 randomly selected mRNAs and three well known thermometers - *rpoH*, *ibpA* and *agsA* sequences from *E. coli*. We also compared the *rpoH* sequences from 81 mesophilic proteobacteria. Although, both *rpoH* and *ibpA*, show a higher rate of melting at their RBS compared with the mean of non-thermometers, contrary to our expectations these higher rates are not significant. Surprisingly, we also do not find any significant differences between *rpoH* thermometers from other γ-proteobacteria and *E. coli* non-thermometers.

4.1 Introduction

Many microorganisms live in a variable environment. They have evolved a variety of mechanisms to sense changes in their environment and alter their gene expression in response to these changes. Regulatory proteins often play a role in controlling the level of transcription and translation of other genes. However, in certain cases post-transcriptional mechanisms, such as changes in mRNA conformation, are known to influence gene expression. In some prokaryotes, reaction to changes in the temperature is thought to be mediated by one such class of mRNAs called RNA thermometers (Yuzawa *et al.*, 1993; Nakahigashi *et al.*, 1995; Morita *et al.*, 1999; Chowdhury *et al.*, 2003; Narberhaus *et al.*, 2006). At lower temperatures, the thermosensing region in these sequences adopts a secondary structure that sequesters the ribosome binding site (RBS) of a gene, hence interfering with translation initiation by the ribosome. At higher temperatures, this thermosensing region upstream of the coding sequence melts, increasing the accessibility of the RBS leading to an increase in the initiation of translation and, in turn, its protein production rate (de Smit and van Duin, 1990; Yuzawa *et al.*, 1993; Chowdhury *et al.*, 2003; Narberhaus *et al.*, 2006).

Previous work on RNA thermometers has focused primarily on understanding and identifying their sequence based features and residues important for thermosensing (de Smit and van Duin, 1990; Yuzawa *et al.*, 1993; Nakahigashi *et al.*, 1995; Morita *et al.*, 1999). Time elapsed spectral studies (Chowdhury *et al.*, 2006) and mutational analyses (Yuzawa *et al.*, 1993; Nakahigashi *et al.*, 1995; Morita *et al.*, 1999) of the thermometer genes have been used to identify regions, which play a crucial part in the thermosensing property. For instance, in one of the most studied RNA thermometer called the ROSE (Repression Of heat-Shock gene Expression) element, a guanine residue at position 83, paired opposite the Shine-Dalgarno (SD) sequence in a hairpin structure is known to play a prominent role in the ability of the mRNA to change its expression with temperature (Chowdhury *et al.*, 2003).

Although these studies provide insights into the mechanisms by which specific thermometers function, little is known as to how widespread these mechanisms are. The fraction of genes in a genome that possess an ability to regulate their translation by thermosensing or a similar mechanism is unknown. More importantly, because the above studies do not include non-thermometers as controls, it is difficult to ascertain if RNA thermometers are a special class of molecules different from other RNAs. Since it is not feasible to perform mutational or spectral studies on every gene to identify whether it behaves as an RNA thermometer, computational tools need to be developed to provide these insights. We here propose a computational approach to characterize RNA thermometers and ask how they differ from non-thermometers in their ability to melt with an increase in temperature. Understanding the melting potential of non-thermometers should aid in understanding the adaptive features of RNA thermometer sequences. We focus specifically on the ability of genes to change their expression by modifying the accessibility of RBS, or in other words, 'RBS exposure'.

Earlier attempts to identify potential RNA thermometers have focused on search patterns based on similarities in the secondary structure of the mRNAs (Waldminghaus *et al.*, 2005, 2007). However, the use of a fixed length sequence for secondary structure limits the utility of this approach. For instance, sequences that differ by only a single nucleotide in their lengths can have drastic differences in their predicted secondary structures (Hughes and McElwaine, 2006). Secondly, most studies when looking at secondary structures of RNAs use mainly the least free energy (LFE) structures. Although, this approach of using the most stable structures has proved useful, there are certain shortcomings when used for characterizing RNA thermometers. It has been shown that as temperature increases, the overall probability and uniqueness of finding a structure in its LFE state decreases (Huynen *et al.*, 1997; Voss *et al.*, 2004). Thus, such an approach could lead to spurious results as the energy landscape of the molecule evolves with temperature (Fig. 4.1). In addition, looking at LFE structures at a single temperature alone provides no means

of quantifying the effect of temperature on the structure. Finally, any pattern-based approach to finding thermometers is restrictive, as it does not take into account novel structures that might be thermosensing.

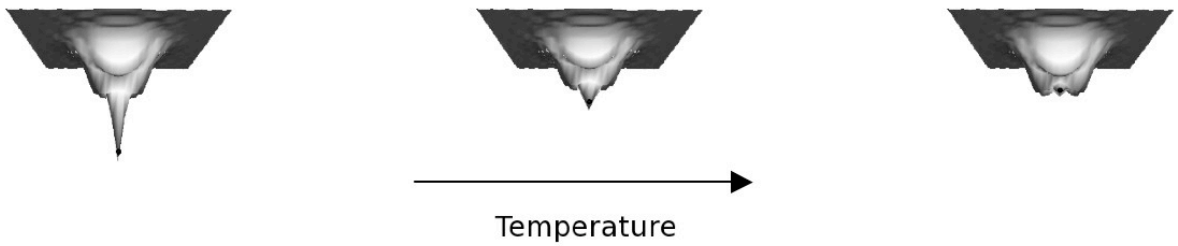


Figure 4.1: Effect of temperature on the energy landscape.
As temperature increases, the probability of finding an mRNA in its most stable state decreases. This is because at higher temperatures, molecules have more energy enabling them to spend more time in higher energy states. Also, at higher temperatures, as the energy landscape becomes flatter, uniqueness of the stable state may also be lost (Huynen *et al.*, 1997).

Here we propose a novel method of quantitatively studying secondary structures of RNAs that addresses all of the above shortcomings. This method explores the ability of mRNAs to change their rate of translation initiation with temperature. We see this approach as complementary to experimental studies in the field of RNA structures.

4.2 Methods

We used the RNAsubopt package from RNA folding software Vienna (Hofacker *et al.*, 1994) to predict secondary structures of the RNAs. This package was used to sample 1000 secondary structures at each temperature for every gene from the entire distribution of structures at that temperature. The sampling of sub-optimal

structures is important because RNA secondary structures with very similar free energies can have drastic differences in their secondary structures (Voss *et al.*, 2004), which might not be captured when looking at the structure with least energy in isolation. The program RNAsubopt generates structures with *probabilities equal to their Boltzmann weights* via stochastic backtracking in the partition function (Wuchty *et al.*, 1999). Since these structures are drawn based on their Boltzmann weights, the entire ensemble of 1000 structures can be viewed as a time ensemble, i.e., the probability of finding a particular structure in our ensemble is proportional to the amount of time the RNA is found to be in that structure. Thus, stable structures would have higher Boltzmann weights and the RNA would spend a greater amount of time in that structure.

In order to understand the effect of temperature on gene expression as measured by RBS exposure, we randomly selected 100 non-thermometer mRNAs from the *E. coli* genome (Shah and Gilchrist, 2010a) as well as *rpoH* mRNA sequence, a known thermometer, from 81 mesophilic γ-proteobacteria for this study (Shah and Gilchrist, 2010a). Transcript start and end positions for *E. coli* genes were obtained from the RegulonDB database (Salgado *et al.*, 2006). Information regarding the position of RBS on the transcript was obtained from the *flexrbs* dataset (Shultzaberger *et al.*, 2001; Shah and Gilchrist, 2010a). We used the entire length of the mRNA (5′ UTR + ORF + 3 UTR) to generate the sub-optimal structures. This was done for the following reasons. The secondary structure of mRNA is highly dependent on the length of the sequence used for simulation (Hughes and McElwaine, 2006). Using a shorter length may prevent detection of any long-range interactions that might be crucial for the stability, and function of the RNA molecule. Moreover, although translation is coupled with transcription in prokaryotes, the half-life of an mRNA is considerably longer than the time required for translation (Bernstein *et al.*, 2002; Hambraeus *et al.*, 2003; Selinger *et al.*, 2003) and hence the mRNA transcript would spend most of its time as a full-length sequence. Thus, we argue that the secondary structure of the mRNA is better simulated by using the

entire mRNA length for our purposes. We also check whether our results are robust to using an mRNA sequence of length 150 nucleotides centered around the RBS. Of the 100 genes from *E. coli*, 56 genes were part of operons. In the case of operons, we simulated the entire mRNA sequence but categorized multiple RBSs within an operon individually.

We simulated 1000 secondary structures of each mRNA at 7 different temperatures ranging from 25 °C to 50 °C. All other parameters in RNAsubopt were used at default values. In order to quantify the openness of RNA, we used a sliding window length of 7 bases to estimate the fraction of simulated structures in which none of the bases in that window were involved in base pairing. A window length of 7 was chosen because the Shine-Dalgarno sequence/RBS in *E. coli* varies from 4-7 bases (Shultzaberger *et al.*, 2001; Kozak, 2005). Changing the window length from 5 bases to 10 bases still resulted in the same qualitative behavior. However, as one would expect, because of the categorical nature of the data (open or close), the fraction of open or melted windows in the structure decreased with window length.

An alternative to sampling structures based on Boltzmann's distribution is to estimate the least free energy (LFE) structures by constraining the RBS in the open conformation (Mathews *et al.*, 2004). The LFE of the constrained and the unconstrained structures can then be used to estimate probability of openness of the RBS. However, as mentioned earlier, with and increase in temperature, the overall probability and uniqueness of finding a structure in its LFE state decreases (Huynen *et al.*, 1997; Voss *et al.*, 2004). Thus, such a method severely limits the ability to compare the probability of openness across temperatures.

In order to compare the probability of openness across temperatures, we fitted a logistic model to the fraction of open windows as a function of temperature.

$$p_i(T) = \frac{e^{a_i + b_i T}}{1 + e^{a_i + b_i T}} \tag{4.1}$$

where $p_i(T)$ is the probability of finding the window at position i in a gene, open at temperature T (°C), a_i and b_i are the intercept and slope parameters of how the log-odds of finding an open window at position i, $log(\frac{p_i(T)}{1-p_i(T)})$, changes with temperature. The ratio $-a_i/b_i$ indicates the temperature at which the probability of openness of a window is 0.5. Although the probability of openness of RBS is positively correlated with protein expression, the exact relationship between the two is unknown.

We find that the logistic model serves as a reasonable descriptor of RNA melting (Fig. 4.2).

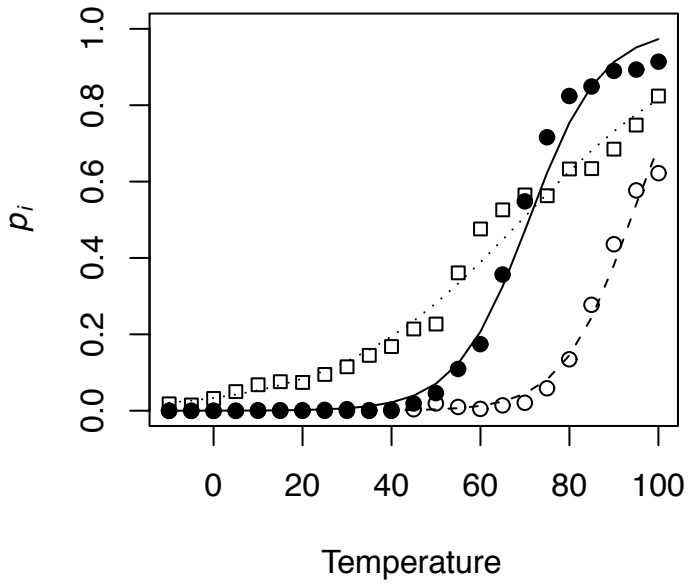


Figure 4.2: Fitting logistic regression.
The solid circles indicate the probability of openness, p_i at the RBS of *rpoH* gene. The open circles and squares represent two randomly chosen windows within *rpoH*. The best fit lines of the logistic regression are given by the solid line for RBS and dashed and dotted line for the randomly chosen windows.

At very low temperatures, we expect most of the bases in the RNA to be paired with other bases. Hence, the probability of openness of a window would approach 0. At very high temperatures, the free energy of base-pairing decreases and most bases would be unpaired causing the probability of openness to approach 1. Thus in a specific range of temperatures, determined by the parameters a and b, we can potentially see a transition between the two states. However, we restrict our simulations to the biological relevant temperature range for mesophiles (25 °C - 55 °C). In this study, we are primarily interested in the parameter b, which describes the rate of change of openness with temperature. For each window within each gene, the Maximum-Likelihood Estimates (MLE) of a and b were calculated using R (R Development Core Team, 2008).

4.3 Results

4.3.1 Capturing the behavior of RNA thermometers

To show that our method is capable of capturing the increase in openness of the RBS of an RNA thermometer, we used the *rpoH* gene sequence of *E. coli*. The *rpoH* gene is a σ-factor involved in the up-regulation of the heat-shock proteins during higher temperatures. It is one of the most studied RNA thermometers (Yuzawa *et al.*, 1993; Nakahigashi *et al.*, 1995; Morita *et al.*, 1999). Fig. 4.3 illustrates how as temperature increases, the RBS of *rpoH* shows a much higher fold-change in openness as compared to the regions flanking it. The openness of the RBS at 50 °C was 25 folds higher than at 25 °C. These results are consistent with the idea that the RBS of a gene might be under stronger selection to increase its openness with temperature.

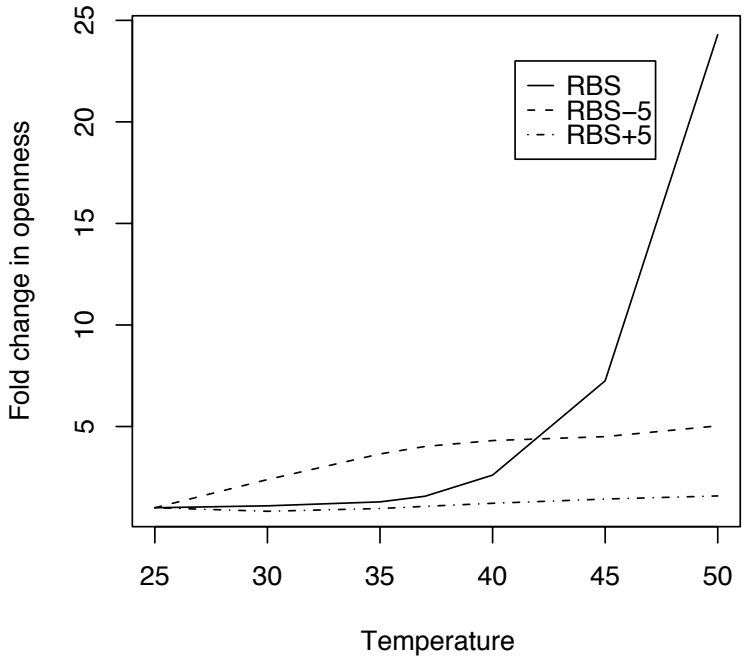


Figure 4.3: Fold-change in the openness of the RBS and regions 5 bases upstream and downstream of it with temperature.
The fold change is with respect to the openness at 25 °C. The RBS of *rpoH* gene has a much higher increase in openness with temperature than the regions around it.

We were also able to replicate the experimental results of Waldminghaus *et al.* (2005) where they showed that the deletion of guanine at position 71 (G71) of the gene *ibpA* in *E. coli*, resulted in a loss of thermosensing activity. Fig. 4.4 shows that both the RNA thermometers *rpoH* and *ibpA* have a higher rate of increase in their RBS exposure compared to the mean of the randomly selected 100 *E. coli* genes. However, the MLE of b drops to 0 when G71 is removed from the *ibpA* gene sequence, as we did not observe a single open window in 1000 runs at all temperatures between 25 °C and 50 °C at that position.

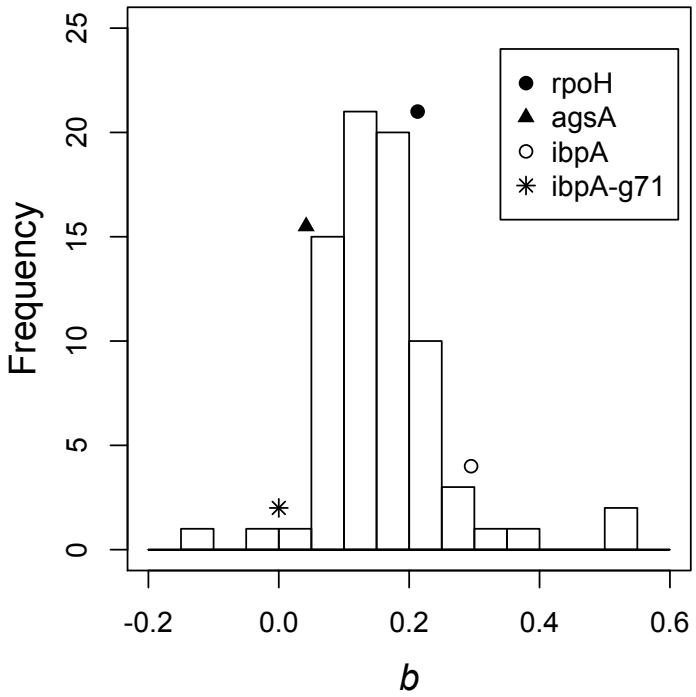


Figure 4.4: The distribution of MLE estimates of b of the 76 genes that differed significantly from zero in *E. coli*.

rpoH, *ibpA* and *agsA* genes show an increase in openness with temperature with b values 0.213, 0.295 and 0.042, respectively. However, none of these values are significantly higher than the mean of the distribution (*Wilcox test*, p-value = 0.156, 0.066 and 0.945, respectively). In addition, when the base G71 is removed from *ibpA* sequence, the MLE estimate of b reduces to 0.

4.3.2 Comparing thermometers and non-thermometers

When the rate of openness of RBS, b, was compared across the 100 genes, we found that b values were not significantly greater than zero for 24 genes at $p\ value = 0.05$. This implies that a small fraction of genes did not show a significant change in openness of its RBS with temperature over the range of temperatures considered. This is surprising because if RNA thermometers were a rare class of mRNAs, then

this number would have been far higher. The distribution of the b values for the remaining 76 genes is shown in Fig. 4.4. Since the distribution of b values is not a Gaussian distribution (*Shapiro-Wilk test*, p-value $< 10^{-5}$), non-parametric tests were employed for further statistical analyses. Although the two of the three RNA thermometers, *rpoH* and *ibpA* had a higher b value than the mean of the entire distribution ($\bar{b} = 0.157$), these higher rates of openness were not significant (*Wilcox test*, p-value $= 0.156$ and p-value $= 0.066$, respectively). Interestingly, we find that RNA thermometer *agsA* had a $b = 0.042$, which, although positive, is lower than the mean of the distribution of b values of non-thermometers. We also show that there is no qualitative difference in our results when considering only 150 nucleotides of the mRNA centered around the RBS (see Figure 4.5).

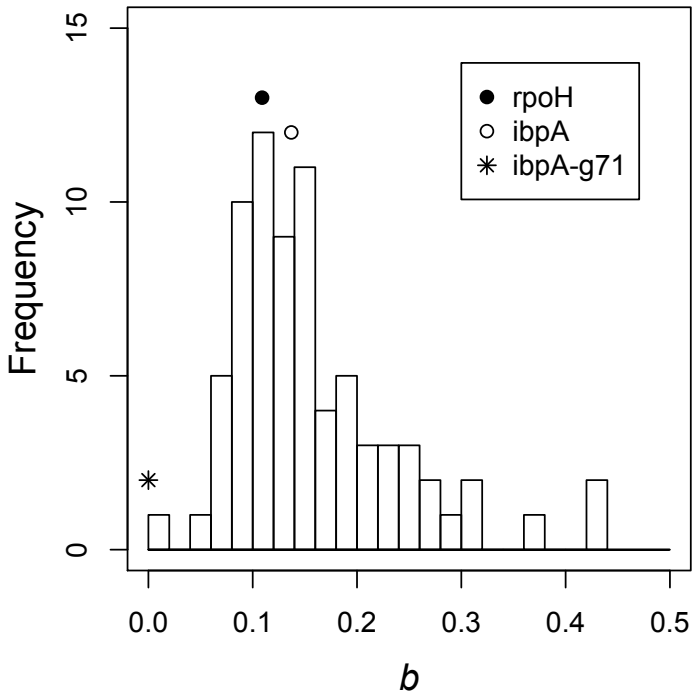


Figure 4.5: The distribution of MLE estimates of b of the 75 genes that differed significantly from zero in *E. coli*.
rpoH, *ibpA* and *agsA* genes show an increase in openness with temperature with b values 0.109, 0.137 and 0.0, respectively. However, none of these values are significantly higher than the mean ($\bar{b} = 0.158$) of the distribution (*Wilcox test*, p-value = 0.781, 0.500 and 0.958, respectively). In addition, when the base G71 is removed from *ibpA* sequence, the MLE estimate of b reduces to 0.

This result did not change even after including non-significant values of b in the above test. This indicates that RNA thermometers do not differ significantly from non-thermometers in increasing the openness of RBS with temperature. It argues that every RNA molecule has an inherent tendency to melt with temperature, albeit to varying degree. These results are also consistent when considering the window

spanning the start codon (ATG) (see Fig. 4.6), stability of which has been shown recently to be correlated with gene expression (Kudla *et al.*, 2009).

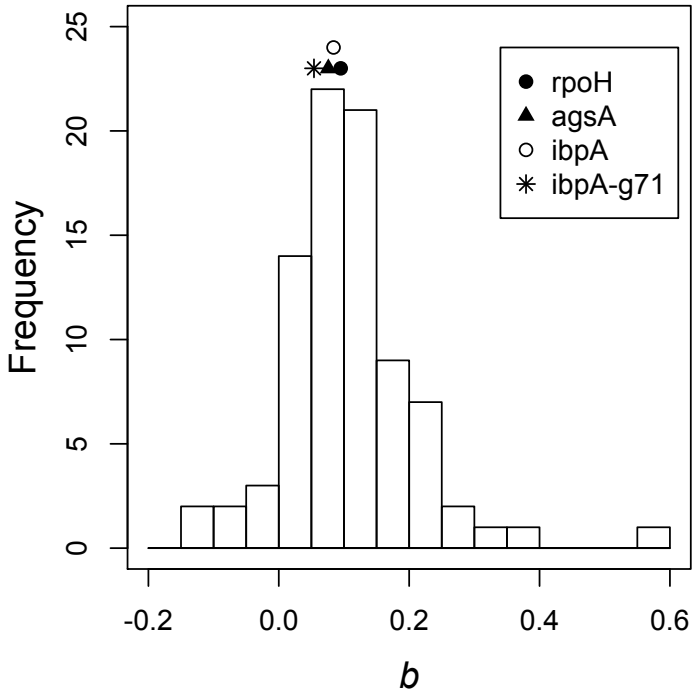


Figure 4.6: The distribution of MLE estimates of b at the start codon (ATG) of the 85 genes that significantly differ from zero in *E. coli*.
rpoH, *ibpA* and *agsA* genes show an increase in openness with temperature with b values 0.095, 0.084 and 0.076, respectively. However, all the values are less than mean of the distribution. Also, when the base G71 is removed from the *ibpA* sequence, the MLE estimate of b reduces to 0.054. The results are consistent with what is observed at the RBS window.

Interestingly, the median transition temperature, given by $-a/b$, was $\sim$ 68 °C. Although the majority of the transition temperatures lie outside the temperature range experienced by mesophiles, it is important to note that this temperature indicates when the probability of openness is 0.5. Although, the relationship between

degree of openness and translation initiation is positively correlated, there exists no quantitative estimate of this relationship. The above values indicate that the RBS needs to be open only a small fraction of time for translation initiation of most genes to meet their target protein production rates.

In order to show the generality of the above results, we compared the distribution of b of *rpoH* of 81 mesophilic γ-proteobacteria to that of the 100 randomly selected genes. Surprisingly, of the 81 *rpoH* sequences, 17 (21%) showed no significant change in their b. We also found that the mean of the two distributions are not significantly different from each other (*Kolmogorov-Smirnov test* p-value = 0.794), further supporting our conclusions. Fig. 4.7 shows the distribution of 76 *E. coli* genes with significant b values alongside the significant b values of *rpoH* genes of 64 mesophilic γ-proteobacteria.

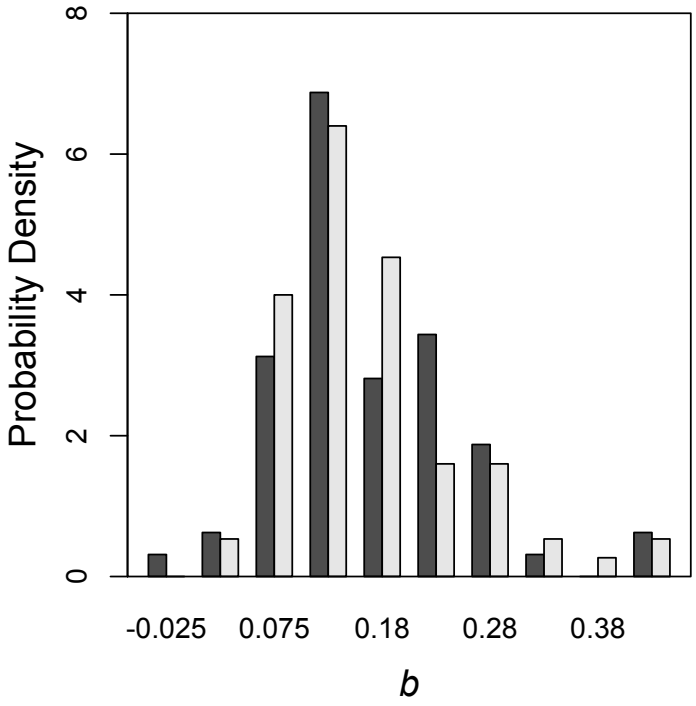


Figure 4.7: Distribution of significant b values of 76 *E. coli* genes and 64 *rpoH* genes of mesophilic γ-proteobacteria.
The two distributions are not significantly different from each other (*Kolmogorov-Smirnov test* p-value $= 0.794$).

4.4 Discussion

We present here a novel method of studying the melting of RNAs with temperature by incorporating the entire distribution of the RNA structures at a given temperature. This approach is more holistic as it takes into account the probability of finding the RNA in a sub-optimal structure based on its free energy as opposed to previous studies which have looked at structures with the least free energies only (Avihoo and Barash, 2005; Morita *et al.*, 1999; Nakahigashi *et al.*, 1995; Waldminghaus *et al.*, 2005, 2007). Although using the minimum free energy structure makes the

analyses of structural features easier, it ignores the sub-optimal, yet highly likely structures that the RNA molecule can also adopt. Using the minimum free energy structure also becomes progressively problematic with and increase in temperature. It has been shown that as temperature increases, probability of finding the RNA in the minimum free energy structure becomes smaller (Huynen *et al.*, 1997) as at higher temperatures, various secondary structures become equally probable as the energy landscape becomes shallower and flatter. Thus, for RNAs whose structure changes with temperature, it becomes important to sample from the entire distribution of structures. In addition, since our approach is not biased towards any particular structural feature, it can be used to identify novel thermosensitive structures.

As one would expect, we find that mRNAs have an inherent tendency to melt with an increase in temperature. This tendency varies with the sequence and the difference in temperatures. Contrary to our expectations, we find that RNA thermometers are not unique with respect to their ability to increase their RBS exposure with temperature. Since it is difficult and expensive to demonstrate the effect of temperature on the RNA secondary structure in the laboratory, researchers have focused primarily on known RNA thermometers. However, due to a lack of such studies on non-thermometers, it has been hard to ascertain whether thermosensing properties are unique to a special class of RNAs. Our results call for further experimental exploration of 'non-thermometers' with changes in temperatures, before firm conclusions can be drawn regarding the uniqueness of RNA thermometers.

Physiological similarities between RNA thermometers and non-thermometers with respect to their melting with temperature, raise an important question that if a large number of mRNAs show an extensive increase in RBS exposure with temperature, why don't we see corresponding changes in their protein expressions. In other words, why do physiological similarities not lead to functional similarities? This discrepancy could be explained, in part, by the fact that the amount of protein expression depends on a variety of factors such as mRNA abundance and stability, amount of regulatory proteins, the stability of the protein itself, and factors apart from the accessibility of

the RBS of the mRNA to the ribosome. Hence, although temperature may not result in significant phenotypic effects of certain genes in terms of protein expression, it does not preclude the possibility of changes in its RBS exposure. Thus, the above results indicate that increased RBS exposure does not solely define as to what constitutes an RNA thermometer.

One of the key challenges in such studies is to devise appropriate measures that quantify the structural features in analyzing the distribution of secondary structures. Here, we use a simple measure of openness to quantify the changes in the structure with temperature. In order to quantify complex structural features like stems and loops in a distribution of RNA structures, more sophisticated measures could be developed. Our analysis based on the current state of RNA folding algorithms is also limited by the simple energy model as well as parameter estimates used in most algorithms.

Another key limitation of this study is the fact that current RNA folding algorithms do not take into account the effect of presence of ribosome on the mRNAs secondary structure. The secondary structure of an mRNA becomes a constantly changing environment due to the presence and movement of ribosomes along the mRNA affecting the openness of a window both upstream and downstream of its current position. Hence, including the effect of ribosomes on the mRNA on translation initiation in the folding algorithm may be important in identifying RNA thermometers computationally. This is likely to be a non-trivial task both mathematically and computationally. However, we believe that incorporating the movement of ribosomes in RNA folding routine would open new avenues of research in investigating and understanding not only the effect of ribosome on the RNA structure and in translation initiation but also on the effect of any RNA-protein interactions on the secondary structure of the RNA.

4.5 Acknowledgments

Funding for this project was provided by the Department of Ecology and Evolutionary Biology at the University of Tennessee, Knoxville. P. S. was additionally supported by National Institute for Mathematical and Biological Synthesis (NIMBioS) Research Assistantship.

Chapter 5

Conclusion

5.1 Synthesis

Over the years, a number of factors have been proposed to explain specific patterns of codon usage bias (CUB). Thus, the problem of understanding these patterns is not of identifying potential evolutionary forces but that of estimating their relative importance (Shah and Gilchrist, 2010b; Plotkin and Kudla, 2011). One of the main challenges in understanding the role played by various selective forces in shaping CUB lies in the fact that there exists no coherent framework to test these hypotheses. The majority of the work on CUB has been correlative and focussed on using heuristic indices to quantify the bias (Bennetzen and Hall, 1982; Sharp and Li, 1987; Wright, 1990). While heuristic approaches play an important role in exploring datasets, especially in the initial stages of analysis, the lack of mechanistic principles sheds little light on cause and effect Moreover, since heuristic indices are based on individual researchers' intuition, they can lead to contradictory results depending on the index used (Stoletzki and Eyre-Walker, 2007; Gilchrist *et al.*, 2009).

In contrast, building upon the insights developed in (Gilchrist and Wagner, 2006; Gilchrist, 2007), we have developed a robust framework of incorporating mechanistic models of protein translation into classical population genetics models to understand CUB. Since the models developed in this work are based on mechanistic principles, observed patterns can be related directly to underlying

biological mechanisms. Thus, we have laid the groundwork upon which mechanistic models of various hypotheses can be simultaneously compared and evaluated.

5.1.1 Consensus and disagreement

While explanations for certain patterns of CUB are generally agreed upon, others are widely debated. For instance, the work presented here as well as other previous studies suggests that CUB in genes with low expression is driven primarily by biased mutation rates (Chamary *et al.*, 2006; Hershberg and Petrov, 2008; Subramanian, 2008). This is due to the fact that in genes expressed at low levels, the efficacy of selection in driving CUB is weak.

However, in genes with high expression, patterns of CUB are thought be driven primarily by natural selection, although the nature of selection is debated. For instance, selection for translation accuracy predicts that codons at sites that are evolutionarily conserved among proteins, will be better at minimizing missense errors than their coding synonyms (Akashi, 1994; Arava *et al.*, 2005; Drummond and Wilke, 2008). This is because, evolutionarily conserved sites are thought to be functionally or structurally important and errors at these sites might render the protein nonfunctional. Preference of codons with high tRNA abundances at these sites is thus thought to support this hypothesis. However, as we show in Chapter 2, the assumption that codons with high tRNA abundances lead to fewer errors is not always true and thus selection for translation accuracy is insufficient in explaining the presence of codons with high tRNA abundance at conserved sites (Shah and Gilchrist, 2010b).

In addition, it has been observed that the codons at the start of a gene are either randomly distributed or have a higher proportion of suboptimal codons that the rest of the sequence. The presence of slow or suboptimal codons at the beginning of a gene is thought to be adaptive for efficient ribosome queueing and prevention of collisions among ribosomes translating a given mRNA (Tuller *et al.*, 2010).

However, (Qin *et al.*, 2004; Gilchrist, 2007; Gilchrist *et al.*, 2009) suggest that the presence of suboptimal codon at the beginning of a gene can also be explained by non-adaptive forces. Selection against nonsense errors predicts that the degree of adaptation in coding sequences should increase along the length of a gene. This is due to the fact that nonsense errors later in the sequence are more energetically expensive than earlier in the sequence as the cell has invested greater resources in making the polypeptide. Since the cost of premature translation termination at the beginning of gene is relatively small, efficacy of selection in maintaining optimal codons may be weak. In a study done with Drs. Michael Gilchrist and Russell Zaretzki, Gilchrist *et al.* (2009) show that this is indeed the case and that the degree of adaptation in codon usage to minimize nonsense errors increases not only along the length of a gene but also with gene expression.

5.2 Beyond translation

Understanding the factors responsible for shaping patterns of codon usage provides important insights and estimates of processes affecting the fundamental process of protein translation. However, insights gained from this understanding has far-reaching implications for a wide range of fields including that of epidemiology, systems biology and organismal and molecular evolution.

5.2.1 Identifying genes under selection

With the exponential growth in genomic data, it is now possible to identify the sets of genes that are under strong selection in various species. Identifying these genes can allow us to make inferences about the organisms's environment as well as on its ecology. For instance, the degree to which an aquatic organism expresses DNA UV repair pathways should reflect the amount of time it spends in the upper reaches of the water column (Buma *et al.*, 2003).

Traditionally the nature of selection acting on a gene - stabilizing or directional, is identified by comparing it with orthologues from its closely related species. The ratio of non-synonymous to synonymous substitutions (dN/dS) in these sequences provides a measure of type of selection the gene is under (Nei and Gojobori, 1986; Yang, 1998). If dN/dS$\ll$ 1, the sequence is thought be under stabilizing or purifying selection and if dN/dS$\gg$ 1, the sequence is thought be under positive or directional selection. However, one of the fundamental assumptions made in this analysis is that synonymous substitutions are neutral. As shown in this work, this is overly simplistic and could lead to various biases. The work presented here allows us to quantify the strength of selection on synonymous codons of a sequence given its expression level and will help in defining better measures of selection.

An alternative to using dN/dS is using heuristic measures of codon usage bias (e.g. RSCU, CAI, F_{op}, E(g), N_c, CBI, CodonO, and RCB (Sharp and Li, 1987; Ikemura, 1981; Karlin and Mrázek, 2000; Wright, 1990; Bennetzen and Hall, 1982; Wan *et al.*, 2006)). As mentioned earlier, a variety of heuristic measures have been developed to quantify the degree of bias in a statistical sense. In contrast to these heuristic measures, we have also developed an index of adaptation based on a specific biological process (Gilchrist *et al.*, 2009). In any case, such measures allow us to identify genes that are under selection using the degree of bias observed in their codon patterns.

5.2.2 Phylogenetic inference and codon bias

One of the fundamental challenges in evolutionary biology is to understand the phylogenetic relationships among organisms. In recent years, molecular data has replaced morphological traits in building phylogenetic trees (Jukes and Cantor, 1969; Fink, 1986; Posada and Crandall, 1998). Models for building phylogenetic trees using gene sequences can be broadly classified into two categories - nucleotide based and codon based models (Goldman and Yang, 1994). As the name

suggests, nucleotide based models account for changes among sequences at the level of individual nucleotides by accounting for heterogeneity in mutations rates among various nucleotides. Codon based models use codons as the fundamental unit of change when building trees rom multiple organisms. In reality, codon based models are really amino acid based models as they account for changes in only those codons that lead to different amino acid. This is generally done by penalizing codon substitutions based on the differences in properties of amino acids that are substituted. As in the case of dN/dS, synonymous substitutions are generally thought to be neutral. In contrast, along with Drs. Laura Kubatko (OSU) and Michael Gilchrist, I have worked on developing codon based models for phylogenetic inference that explicitly takes into account the effects of synonymous substitutions. Such models would potentially provide greater resolution and lead to more accurate phylogenies.

5.2.3 Codon usage and medicine

A large number of sequenced organisms are pathogens.However, it is unlikely that our understanding of these organisms is ever going to rival that of model organisms. For many of them, their sequence data might be the only source of information we may have for a while. Thus by parsing genomic patterns such as those of codon usage in an evolutionary context can help us understand the biology of the organism. For example, it has been shown that the patterns of codon usage in many viruses reflect an adaptation to the tRNA pools of their host (Zhou *et al.*, 1999; Plotkin and Dushoff, 2003; Grote *et al.*, 2005; Coleman *et al.*, 2008). Recently, (Coleman *et al.*, 2008) showed that by changing only the codon usage of a virus genome but keeping the amino acid sequence same, one can dramatically reduce the infectivity of the virus. Moreover, since the virus still produces the same proteins, albeit at a much lower rate, it elicits the same immune response and thus such modified viruses could be used for developing vaccines.

Bibliography

Bibliography

Agris, P. F., 1991 Wobble position modified nucleosides evolved to select transfer RNA codon recognition: a modified-wobble hypothesis. Biochimie **73**: 1345–9. 37

Agris, P. F., F. A. P. Vendeix, and W. D. Graham, 2007 tRNA's wobble decoding of the genome: 40 years of modification. J. Mol. Biol. **366**: 1–13. 37

Akashi, H., 1994 Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. Genetics **136**: 927–35. 10, 19, 32, 34, 46, 67, 96

Akashi, H., 2001 Gene expression and molecular evolution. Curr. Opin. Genet. Devel. **11**: 660–6. 10

Akashi, H., 2003 Translational selection and yeast proteome evolution. Genetics **164**: 1291–303. 47

Akashi, H., and A. Eyre-Walker, 1998 Translational selection and molecular evolution. Curr. Opin. Genet. Devel. **8**: 688–93. 46, 47

Akashi, H., and T. Gojobori, 2002 Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. Proc. Natl. Acad. Sci. U.S.A. **99**: 3695–700. 3, 10

Alberts, B., A. Johnson, J. Lewis, M. Raff, K. Roberts, *et al.*, 2008 *Molecular Biology of the Cell. 5th Edition*. Garland Science, New York, NY. 47

Andersson, D. I., K. Bohman, L. A. Isaksson, and C. G. Kurland, 1982 Translation rates and misreading characteristics of rpsd mutants in *Escherichia coli*. Mol Gen Genet **187**: 467–72. 11, 38

Arava, Y., F. E. Boas, P. O. Brown, and D. Herschlag, 2005 Dissecting eukaryotic translation and its control by ribosome density mapping. Nucleic Acids Res. **33**: 2421–32. 10, 29, 34, 46, 96

Arava, Y., Y. Wang, J. D. Storey, C. L. Liu, P. O. Brown, *et al.*, 2003 Genome-wide analysis of mRNA translation profiles in saccharomyces cerevisiae. Proc. Natl. Acad. Sci. U.S.A. **100**: 3889–94. 49

Ardell, D. H., and G. Sella, 2001 On the evolution of redundancy in genetic codes. J. Mol. Evol. **53**: 269–81. 36

Avihoo, A., and D. Barash, 2005 Temperature and mutation switches in the secondary structure of small RNAs. IEEE Computational Systems Bioinformatics Conference, 2005. Workshops and Poster Abstracts : 235–236. 91

Bennetzen, J. L., and B. D. Hall, 1982 Codon selection in yeast. J. Biol. Chem. **257**: 3026–31. 5, 46, 66, 95, 98

Berg, J., S. Willmann, and M. Lässig, 2004 Adaptive evolution of transcription factor binding sites. BMC Evol. Biol. **4**: 42. 71

Berg, O. G., and C. G. Kurland, 1997 Growth rate-optimised tRNA abundance and codon usage. J. Mol. Biol. **270**: 544–50. 10

Bernstein, J. A., A. B. Khodursky, P.-H. Lin, S. Lin-Chao, and S. N. Cohen, 2002 Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays. Proc. Natl. Acad. Sci. U.S.A. **99**: 9697–702. 81

Beyer, A., J. Hollunder, H.-P. Nasheuer, and T. Wilhelm, 2004 Post-transcriptional expression regulation in the yeast saccharomyces cerevisiae on a genomic scale. Mol. Cell Proteomics **3**: 1083–92. 49, 64

Blanchard, S. C., R. L. Gonzalez, H. D. Kim, S. Chu, and J. D. Puglisi, 2004a tRNA selection and kinetic proofreading in translation. Nat. Struct. Mol. Biol. **11**: 1008–14. 37

Blanchard, S. C., H. D. Kim, R. L. Gonzalez, J. D. Puglisi, and S. Chu, 2004b tRNA dynamics on the ribosome during translation. Proc. Natl. Acad. Sci. U.S.A. **101**: 12893–8. 37

Bouadloun, F., D. Donner, and C. G. Kurland, 1983 Codon-specific missense errors in vivo. EMBO J. **2**: 1351–6. 11

Bulmer, M., 1988 Are codon usage patterns in unicellular organisms determined by selection-mutation balance? J. Evol. Biol. **1**: 15–26. 58

Bulmer, M., 1991 The selection-mutation-drift theory of synonymous codon usage. Genetics **129**: 897–907. 3, 5, 6, 10, 34, 46, 47, 48

Buma, A., P. Boelen, and W. Jeffrey, 2003 Uvr-induced DNA damage in aquatic organisms. UV effects in aquatic organisms and ecosystems : 291–327. 97

Carbone, A., A. Zinovyev, and F. Képès, 2003 Codon adaptation index as a measure of dominating codon bias. Bioinform. **19**: 2005–15. 2

Chamary, J. V., J. L. Parmley, and L. D. Hurst, 2006 Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat. Rev. Genet. **7**: 98–108. 3, 96

Chan, P. P., and T. M. Lowe, 2009 GtRNAdb: a database of transfer RNA genes detected in genomic sequence. Nucleic Acids Res. **37**: D93–7. 12, 70

Chowdhury, S., C. Maris, F. H.-T. Allain, and F. Narberhaus, 2006 Molecular basis for temperature sensing by an RNA thermometer. EMBO J. **25**: 2487–2497. 78

Chowdhury, S., C. Ragaz, E. Kreuger, and F. Narberhaus, 2003 Temperature-controlled structural alterations of an RNA thermometer. J. Biol. Chem. **278**: 47915–21. 6, 78

Cognat, V., J.-M. Deragon, E. Vinogradova, T. Salinas, C. Remacle, *et al.*, 2008 On the evolution and expression of *Chlamydomonas reinhardtii* nucleus-encoded transfer RNA genes. Genetics **179**: 113–23. 12, 36

Coleman, J. R., D. Papamichail, S. Skiena, B. Futcher, E. Wimmer, *et al.*, 2008 Virus attenuation by genome-scale changes in codon pair bias. Science **320**: 1784–7. 46, 99

Cornut, B., and R. C. Willson, 1991 Measurement of translational accuracy in vivo: missense reporting using inactive enzyme mutants. Biochimie **73**: 1567–72. 4

Curran, J. F., and M. Yarus, 1989 Rates of aminoacyl-tRNA selection at 29 sense codons in vivo. J. Mol. Biol. **209**: 65–77. 20, 37, 70

de Smit, M. H., and J. van Duin, 1990 Secondary structure of the ribosome binding site determines translational efficiency: a quantitative analysis. Proc. Natl. Acad. Sci. U.S.A. **87**: 7668–72. 6, 78

Dobzhansky, T., 1973 Nothing in biology makes sense except in the light of evolution. Am Biol Teach **35**: 125–129. 1

Dong, H., L. Nilsson, and C. G. Kurland, 1996 Co-variation of tRNA abundance and codon usage in *Escherichia coli* at different growth rates. J. Mol. Biol. **260**: 649–63. 2, 12, 26, 36, 46, 70

dos Reis, M., R. Savva, and L. Wernisch, 2004 Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res. **32**: 5036–44. 46

Drummond, D. A., J. D. Bloom, C. Adami, C. O. Wilke, and F. H. Arnold, 2005 Why highly expressed proteins evolve slowly. Proc. Natl. Acad. Sci. U.S.A. **102**: 14338–43. 34, 35, 67

Drummond, D. A., A. Raval, and C. O. Wilke, 2006 A single determinant dominates the rate of yeast protein evolution. Mol. Biol. Evol. **23**: 327–37. 34

Drummond, D. A., and C. O. Wilke, 2008 Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell **134**: 341–52. 32, 67, 96

Drummond, D. A., and C. O. Wilke, 2009 The evolutionary consequences of erroneous protein synthesis. Nat. Rev. Genet. **10**: 715–24. 10, 11, 19, 34, 46, 68

Duret, L., and D. Mouchiroud, 1999 Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. Proc. Natl. Acad. Sci. U.S.A. **96**: 4482–7. 58

Eyre-Walker, A., 1996 Synonymous codon bias is related to gene length in *Escherichia coli*: selection for translational accuracy? Mol. Biol. Evol. **13**: 864–72. 3

Fink, W. L., 1986 Microcomputers and phylogenetic analysis. Science **234**: 1135–9. 98

Fitch, W. M., 1976 Is there selection against wobble in codon-anticodon pairing? Science **194**: 1173–4. 1

Fluitt, A., E. Pienaar, and H. Viljoen, 2007 Ribosome kinetics and aa-tRNA competition determine rate and fidelity of peptide synthesis. Comput. Biol. Chem. **31**: 335–46. 4, 11, 20, 33, 37

Freeland, S. J., and L. D. Hurst, 1998 The genetic code is one in a million. J. Mol. Evol. **47**: 238–48. 11, 35

Freeland, S. J., R. D. Knight, L. F. Landweber, and L. D. Hurst, 2000 Early fixation of an optimal genetic code. Mol. Biol. Evol. **17**: 511–8. 11

Gavrilets, S., 2004 *Fitness landscapes and the origin of species: monographs in population biology Vol. 41*. Princeton Univ. Press, Princeton, NJ. 48, 71

Gilchrist, M. A., 2007 Combining models of protein translation and population genetics to predict protein production rates from codon usage patterns. Mol. Biol. Evol. **24**: 2362–72. 5, 33, 34, 46, 47, 48, 49, 64, 70, 95, 97

Gilchrist, M. A., P. Shah, and R. Zaretzki, 2009 Measuring and detecting molecular adaptation in codon usage against nonsense errors during protein translation. Genetics **183**: 1493–505. 10, 33, 34, 35, 46, 47, 48, 68, 95, 97, 98

Gilchrist, M. A., and A. Wagner, 2006 A model of protein translation including codon bias, nonsense errors, and ribosome recycling. J. Theor. Biol. **239**: 417–34. 3, 34, 46, 68, 70, 95

Goldman, N., and Z. Yang, 1994 A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. **11**: 725–36. 98

Grantham, R., 1974 Amino acid difference formula to help explain protein evolution. Science **185**: 862–4. 11

Grantham, R., C. Gautier, and M. Gouy, 1980 Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type. Nucleic Acids Res. **8**: 1893–912. 1

Greenbaum, D., C. Colangelo, K. Williams, and M. Gerstein, 2003 Comparing protein abundance and mRNA expression levels on a genomic scale. Genome Biol. **4**: 117. 35

Gromadski, K. B., and M. V. Rodnina, 2004 Kinetic determinants of high-fidelity tRNA discrimination on the ribosome. Mol. Cell. **13**: 191–200. 4, 11, 20, 37, 66

Grote, A., K. Hiller, M. Scheer, R. Münch, B. Nörtemann, *et al.*, 2005 JCat: a novel tool to adapt codon usage of a target gene to its potential expression host. Nucl Acids Res **33**: W526–31. 99

Guo, H. H., J. Choe, and L. A. Loeb, 2004 Protein tolerance to random amino acid change. Proc. Natl. Acad. Sci. U.S.A. **101**: 9205–10. 4, 67, 68

Hambraeus, G., C. V. Wachenfeldt, and L. Hederstedt, 2003 Genome-wide survey of mRNA half-lives in *Bacillus subtilis* identifies extremely stable mRNAs. Mol. Genet. Genomics **269**: 706–714. 81

Hershberg, R., and D. A. Petrov, 2008 Selection on codon bias. Annu Rev Genet **42**: 287–99. 47, 96

Higgs, P., 2009 A four-column theory for the origin of the genetic code: tracing the evolutionary pathways that gave rise to an optimized code. Biol Direct **4**: 16. 11

Hofacker, I., W. Fontana, and P. Stadler, 1994 Fast folding and comparison of RNA secondary structures. Monatshefte für Chemie **125**: 167–188. 80

Hughes, T., and J. McElwaine, 2006 Mathematical and biological modelling of RNA secondary structure and its effects on gene expression. Comput. Math. Methods Med. **7**: 37–43. 79, 81

Huynen, M., R. Gutell, and D. Konings, 1997 Assessing the reliability of RNA folding using statistical mechanics. J. Mol. Biol. **267**: 1104–12. 79, 80, 82, 92

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. J. Mol. Biol. **151**: 389–409. 1, 2, 4, 5, 32, 46, 98

Ikemura, T., 1985 Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. **2**: 13–34. 19, 35

Jørgensen, F., and C. G. Kurland, 1990 Processivity errors of gene expression in *Escherichia coli*. J. Mol. Biol. **215**: 511–21. 4, 11, 38

Jukes, T. H., and C. R. Cantor, 1969 *Evolution of protein molecules*. Academic Press, NewYork. 98

Kanaya, S., Y. Yamada, Y. Kudo, and T. Ikemura, 1999 Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene **238**: 143–55. 2, 10, 12, 36, 70

Karlin, S., and J. Mrázek, 2000 Predicted highly expressed genes of diverse prokaryotic genomes. J. Bacteriol. **182**: 5238–50. 98

Kellogg, E., and N. Juliano, 1997 The structure and function of RuBisCO and their implications for systematic studies. Am. J. Bot. . 35

Kimchi-Sarfaty, C., J. M. Oh, I.-W. Kim, Z. E. Sauna, A. M. Calcagno, *et al.*, 2007 A "silent" polymorphism in the mdr1 gene changes substrate specificity. Science **315**: 525–8. 34, 46

Kimura, M., 1964 Diffusion models in population genetics. Journal of Applied Probability **1**. 71

Kothe, U., and M. V. Rodnina, 2007 Codon reading by tRNA-ala with modified uridine in the wobble position. Mol. Cell. **25**: 167–74. 37

Kozak, M., 2005 Regulation of translation via mRNA structure in prokaryotes and eukaryotes. Gene **361**: 13–37. 82

Kramer, E. B., and P. J. Farabaugh, 2007 The frequency of translational misreading errors in *e. coli* is largely determined by tRNA competition. RNA **13**: 87–96. 4, 11, 19, 25, 32, 33, 36

Kudla, G., A. W. Murray, D. Tollervey, and J. B. Plotkin, 2009 Coding-sequence determinants of gene expression in *Escherichia coli*. Science **324**: 255–8. 6, 46, 68, 89

Kurland, C., 1987 Strategies for efficiency and accuracy in gene expression. Trends Biochem. Sci. **12**: 126–128. 47

Kurland, C., and J. Gallant, 1996 Errors of heterologous protein expression. Curr Opin Biotechnol **7**: 489–93. 3, 4

Kurland, C. G., 1992 Translational accuracy and the fitness of bacteria. Annu Rev Genet **26**: 29–50. 3

Kurland, C. G., and M. Ehrenberg, 1987 Growth-optimizing accuracy of gene expression. Annual review of biophysics and biophysical chemistry **16**: 291–317. 11

Lee, J. W., K. Beebe, L. A. Nangle, J. Jang, C. M. Longo-Guess, *et al.*, 2006 Editing-defective tRNA synthetase causes protein misfolding and neurodegeneration. Nature **443**: 50–5. 4

Lim, V. I., and J. F. Curran, 2001 Analysis of codon:anticodon interactions within the ribosome provides new insights into codon reading and the genetic code structure. RNA **7**: 942–57. 17, 37, 70

Lobley, G. E., V. Milne, J. M. Lovie, P. J. Reeds, and K. Pennie, 1980 Whole body and tissue protein synthesis in cattle. Br J Nutr **43**: 491–502. 10

Lovmar, M., and M. Ehrenberg, 2006 Rate, accuracy and cost of ribosomes in bacterial cells. Biochimie **88**: 951–61. 47

MacKay, V. L., X. Li, M. R. Flory, E. Turcott, G. L. Law, *et al.*, 2004 Gene expression analyzed by high-resolution state array analysis and quantitative proteomics: response of yeast to mating pheromone. Mol. Cell Proteomics **3**: 478–89. 49

Manley, J. L., 1978 Synthesis and degradation of termination and premature-termination fragments of β-galactosidase in vitro and in vivo. J. Mol. Biol. **125**: 407–32. 4

Marin, M., 2008 Folding at the rhythm of the rare codon beat. Biotechnol J **3**: 1047–57. 34

Markiewicz, P., L. G. Kleina, C. Cruz, S. Ehret, and J. H. Miller, 1994 Genetic studies of the lac repressor. XIV. analysis of 4000 altered *Escherichia coli* lac repressors reveals essential and non-essential residues, as well as "spacers" which do not require a specific sequence. J. Mol. Biol. **240**: 421–33. 4, 67

Mathews, D. H., M. D. Disney, J. L. Childs, S. J. Schroeder, M. Zuker, *et al.*, 2004 Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. Proc. Natl. Acad. Sci. U.S.A. **101**: 7287–92. 82

Menninger, J. R., 1977 Ribosome editing and the error catastrophe hypothesis of cellular aging. Mech Ageing Dev **6**: 131–142. 3

Morita, M. T., Y. Tanaka, T. S. Kodama, Y. Kyogoku, H. Yanagi, *et al.*, 1999 Translational induction of heat shock transcription factor σ-32: evidence for a built-in rna thermosensor. Genes Devel. **13**: 655–65. 6, 78, 84, 91

Mougel, F., C. Manichanh, G. D. N'guyen, and M. Termier, 2004 Genomic choice of codons in 16 microbial species. J. Biomol. Struct. Dyn. **22**: 315–29. 2

Musto, H., H. Romero, and A. Zavala, 2003 Translational selection is operative for synonymous codon usage in *Clostridium perfringens* and *Clostridium acetobutylicum*. Microbiology **149**: 855–63. 58

Nakahigashi, K., H. Yanagi, and T. Yura, 1995 Isolation and sequence analysis of rpoh genes encoding σ-32 homologs from gram negative bacteria: conserved mRNA and protein segments for heat shock regulation. Nucleic Acids Res. **23**: 4383–90. 6, 78, 84, 91

Narberhaus, F., T. Waldminghaus, and S. Chowdhury, 2006 RNA thermometers. FEMS Microbiol. Rev. **30**: 3–16. 6, 78

Nei, M., and T. Gojobori, 1986 Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3**: 418–26. 98

Ogle, J. M., D. E. Brodersen, W. M. Clemons, M. J. Tarry, A. P. Carter, *et al.*, 2001 Recognition of cognate transfer RNA by the 30S ribosomal subunit. Science **292**: 897–902. 11

Ogle, J. M., and V. Ramakrishnan, 2005 Structural insights into translational fidelity. Annu Rev Biochem **74**: 129–77. 4

Pannevis, M. C., and D. F. Houlihan, 1992 The energetic cost of protein synthesis in isolated hepatocytes of rainbow trout (oncorhynchus mykiss). J Comp Physiol B, Biochem Syst Environ Physiol **162**: 393–400. 10

Peixoto, L., V. Fernandez, and H. Musto, 2004 The effect of expression levels on codon usage in *Plasmodium falciparum*. Parasitology **128**: 245–251. 58

Plotkin, J. B., and J. Dushoff, 2003 Codon bias and frequency-dependent selection on the hemagglutinin epitopes of influenza a virus. Proc. Natl. Acad. Sci. U.S.A. **100**: 7152–7. 99

Plotkin, J. B., and G. Kudla, 2011 Synonymous but not the same: the causes and consequences of codon bias. Nat. Rev. Genet. **12**: 32–42. 48, 68, 95

Posada, D., and K. A. Crandall, 1998 MODELTEST: testing the model of DNA substitution. Bioinform. **14**: 817–8. 98

Powell, M., 2006 The NEWUOA software for unconstrained optimization without derivatives. Large-Scale Nonlinear Optimization : 255–297. 69

Precup, J., and J. Parker, 1987 Missense misreading of asparagine codons as a function of codon identity and context. J. Biol. Chem. **262**: 11351–5. 11, 29

Qin, H., W. B. Wu, J. M. Comeron, M. Kreitman, and W.-H. Li, 2004 Intragenic spatial patterns of codon usage bias in prokaryotic and eukaryotic genomes. Genetics **168**: 2245–60. 68, 97

R Development Core Team, 2008 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. 84

Rocha, E. P. C., 2004 Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. Genome Res. **14**: 2279–86. 10

Rodnina, M. V., and W. Wintermeyer, 2001 Ribosome fidelity: tRNA discrimination, proofreading and induced fit. Trends Biochem Sci **26**: 124–30. 4

Salgado, H., S. Gama-Castro, M. Peralta-Gil, E. Diaz-Peredo, F. Sanchez-Solano, *et al.*, 2006 Regulondb (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization, and growth conditions. Nucleic Acids Res. **34**: D394–D397. 81

Selinger, D. W., R. M. Saxena, K. J. Cheung, G. M. Church, and C. Rosenow, 2003 Global RNA half-life analysis in *Escherichia coli* reveals positional patterns of transcript degradation. Genome Res. **13**: 216–23. 81

Sella, G., and A. E. Hirsh, 2005 The application of statistical physics to evolutionary biology. Proc. Natl. Acad. Sci. U.S.A. **102**: 9541–6. 48, 49, 66, 71, 72

Shah, P., and M. Gilchrist, 2010a Is thermosensing property of rna thermometers unique? PLoS ONE . 81

Shah, P., and M. A. Gilchrist, 2010b Effect of correlated tRNA abundances on translation errors and evolution of codon usage bias. PLoS Genet. **6**: e1001128. 4, 46, 49, 66, 67, 68, 95, 96

Sharp, P. M., and K. M. Devine, 1989 Codon usage and gene expression level in *Dictyostelium discoideum*: highly expressed genes do 'prefer' optimal codons. Nucleic Acids Res. **17**: 5029–39. 58

Sharp, P. M., and W. H. Li, 1986 An evolutionary perspective on synonymous codon usage in unicellular organisms. J. Mol. Evol. **24**: 28–38. 5, 10, 46, 58

Sharp, P. M., and W. H. Li, 1987 The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. **15**: 1281–95. 95, 98

Shultzaberger, R. K., R. E. Bucheimer, K. E. Rudd, and T. D. Schneider, 2001 Anatomy of *Escherichia coli* ribosome binding sites. J. Mol. Biol. **313**: 215–28. 81, 82

Sørensen, M. A., C. G. Kurland, and S. Pedersen, 1989 Codon usage determines translation rate in *Escherichia coli*. J. Mol. Biol. **207**: 365–77. 38

Stoletzki, N., and A. Eyre-Walker, 2007 Synonymous codon usage in *Escherichia coli*: selection for translational accuracy. Mol. Biol. Evol. **24**: 374–81. 10, 32, 34, 95

Subramanian, S., 2008 Nearly neutrality and the evolution of codon usage bias in eukaryotic genomes. Genetics **178**: 2429–32. 2, 96

Tsai, C.-J., Z. E. Sauna, C. Kimchi-Sarfaty, S. V. Ambudkar, M. M. Gottesman, *et al.*, 2008 Synonymous mutations and ribosome stalling can lead to altered folding pathways and distinct minima. J. Mol. Biol. **383**: 281–91. 34

Tsung, K., S. Inouye, and M. Inouye, 1989 Factors affecting the efficiency of protein synthesis in *Escherichia coli*. production of a polypeptide of more than 6000 amino acid residues. J. Biol. Chem. **264**: 4428–33. 4

Tuller, T., Y. Y. Waldman, M. Kupiec, and E. Ruppin, 2010 Translation efficiency is determined by both codon bias and folding energy. Proc. Natl. Acad. Sci. U.S.A. **107**: 3645–50. 6, 46, 68, 96

Varenne, S., J. Buc, R. Lloubes, and C. Lazdunski, 1984 Translation is a non-uniform process. effect of tRNA availability on the rate of elongation of nascent polypeptide chains. J. Mol. Biol. **180**: 549–76. 4, 11, 32, 33, 38

Vetsigian, K., and N. Goldenfeld, 2009 Genome rhetoric and the emergence of compositional bias. Proc. Natl. Acad. Sci. U.S.A. **106**: 215–20. 36

Voss, B., C. Meyer, and R. Giegerich, 2004 Evaluating the predictability of conformational switching in RNA. Bioinform. **20**: 1573–82. 79, 81, 82

Wagner, A., 2005 Energy constraints on the evolution of gene expression. Mol. Biol. Evol. **22**: 1365–74. 47, 49

Waldminghaus, T., A. Fippinger, J. Alfsmann, and F. Narberhaus, 2005 RNA thermometers are common in α- and γ-proteobacteria. Biol. Chem. **386**: 1279–1286. 79, 85, 91

Waldminghaus, T., L. C. Gaubig, and F. Narberhaus, 2007 Genome-wide bioinformatic prediction and experimental evaluation of potential RNA thermometers. Mol. Genet. Genomics **278**: 555–564. 79, 91

Wan, X., J. Zhou, and D. Xu, 2006 CodonO: a new informatics method for measuring synonymous codon usage bias within and across genomes. Int J Gen Syst **35**: 109–125. 98

Warner, J. R., 1999 The economics of ribosome biosynthesis in yeast. Trends Biochem Sci **24**: 437–40. 3, 5, 10, 47

Wilke, C. O., and D. A. Drummond, 2006 Population genetics of translational robustness. Genetics **173**: 473–81. 35

Wintermeyer, W., F. Peske, M. Beringer, K. B. Gromadski, A. Savelsbergh, *et al.*, 2004 Mechanisms of elongation on the ribosome: dynamics of a macromolecular machine. Biochem. Soc. Trans. **32**: 733–7. 4

Wong, J. T., 1975 A co-evolution theory of the genetic code. Proc. Natl. Acad. Sci. U.S.A. **72**: 1909–12. 36

Wright, F., 1990 The 'effective number of codons' used in a gene. Gene **87**: 23–9. 95, 98

Wright, S., 1969 *Evolution of the genetics of population Volume 2: The theory of gene frequencies*. University of Chicago Press, Chicago. 48

Wuchty, S., W. Fontana, I. L. Hofacker, and P. Schuster, 1999 Complete suboptimal folding of RNA and the stability of secondary structures. Biopolymers **49**: 145–65. 81

YANG, Z., 1998 Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol. Biol. Evol. **15**: 568–73. 98

YUZAWA, H., H. NAGAI, H. MORI, and T. YURA, 1993 Heat induction of σ-32 synthesis mediated by mRNA secondary structure: a primary step of the heat shock response in escherichia coli. Nucleic Acids Res. **21**: 5449–55. 6, 78, 84

ZAHER, H. S., and R. GREEN, 2009 Fidelity at the molecular level: lessons from protein synthesis. Cell **136**: 746–62. 4, 11

ZHAO, L., C. LONGO-GUESS, B. S. HARRIS, J.-W. LEE, and S. L. ACKERMAN, 2005 Protein accumulation and neurodegeneration in the woozy mutant mouse is caused by disruption of SIL1, a cochaperone of BiP. Nat. Genet. **37**: 974–9. 4

ZHOU, J., W. J. LIU, S. W. PENG, X. Y. SUN, and I. FRAZER, 1999 Papillomavirus capsid protein expression level depends on the match between codon usage and tRNA availability. J Virol **73**: 4972–82. 99

Vita

Premal Shah was born in Akola, India on October 23, 1984. He was raised in Chennai, India where he graduated high school from GSSJV in 2002. In 2006 he graduated from the Anna University at Chennai with a B. Tech. in Industrial Biotechnology. He joined the University of Tennessee, Knoxville as a graduate student in the Fall of 2006 and received a Ph. D. in Ecology and Evolutionary Biology in 2011. Premal will begin working as a postdoctoral research fellow in the Department of Biology at University of Pennsylvania, Philadelphia, PA in June 2011.