CODON USAGE BIAS, tRNA MODIFICATIONS AND TRANSLATIONAL SELECTION IN BACTERIA

By

Wenqi Ran

A Thesis

Submitted to the School of Graduate Studies

in Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

McMaster University

DOCTOR OF PHILOSOPHY (2010) McMaster University

(Physics) Hamilton, Ontario

TITLE: Codon Usage, tRNA Modifications and Translational Selection in Bacteria

AUTHOR: Wenqi Ran (Msc. Peking University)

SUPERVISOR: Professor Paul G Higgs

NUMBER OF PAGES: XVI, 139

ABSTRACT

In this thesis, codon usage bias is studied in a large number of bacterial genomes. The codon frequency of each codon is measured as a fraction of the total number of codons for each amino acid, which can be used to calculate the potential selection strength in a genome according to the population genetics theory of selection, mutation and drift. It is assumed that codon frequencies in low expression genes are affected principally by mutation rates, but the frequencies in high expression genes are controlled by both selection and mutation. The ribosomal protein genes and elongation factors are used as high expression genes, while the rest as low expression genes. By comparing the frequencies of codons in high and low expression genes, we can determine the strength of selection acting on high expression genes. A model of translation kinetics is developed, which predicts the way that the speed of translation of each codon depends on the number of copies of each type of tRNA gene in the genome. This theory reveals that codon usage and tRNA gene copy number have to co-adapt to each other to optimize the speed of translation. We show that there are often multiple possible stable combinations of tRNAs and codon usage. This explains the observation that different codons are sometimes preferred for different amino acids in the same organism.

We compare our theory with observed codon frequencies in a large number of bacterial genomes. Codon families are grouped according to different combinations of tRNAs and the averages of codon frequencies for all combinations are taken respectively. A preferred codon is defined as the codon whose frequency increases from low

expression genes to high expression genes. The interesting phenomenon is that the anticodon-codon interactions are different between two-codon families (the U+C and A+G codon families) and four-codon families, mainly because of the tRNA modifications on position 34 of the tRNA anticodon loop. We show that the preferred codon depends on which base is present at wobble position and whether a base modification happens. It is found that the preferred anticodon-codon combinations do not always correspond to Watson-Crick pairs. In particular, in four-codon families, tRNAs with U at the wobble position appear to interact surprisingly well with U-ending codons.

We introduce a parameter, K, which measures the strength of codon bias in each genome. We show that K is strongly correlated with the growth rate of the organism. This is consistent with the idea that selection for translational speed is most important for rapidly multiplying organisms. From the theory, the effective population size N_e is also expected to influence the selection strength. However, our data analysis shows that K is not correlated with $N_e u$. This shows that variation of $N_e u$ among species is not a confounding factor in our interpretation.

Although we believe that translation speed is the main reason for selection on codon usage in bacteria, it is also possible that selection for translational accuracy plays a role. Selection for accuracy can be tested by comparing codon frequencies in conserved and variable sites within the same gene sequences. We introduce a new statistical test to measure the strength of this effect. We observe a small but significant effect of accuracy with this method.

ACKNOWLEDGEMENTS

I would like to express my great thanks to my supervisor, Dr. Paul G Higgs, for his excellent guidance, encouragement and endless support through the past four years. His ambition and passion for science, critical thinking, scientific integrity, breadth of knowledge and hard working attitude impressed me so much. Working in Dr. Higgs' group has been one of my most valuable experiences in my life. He teaches me everything he knows and encourages me to learn everything we don't. He also gave me the great opportunities to all kinds of 1st class conferences in Europe and America, which broaden my view, encourage my passion in science and make me work smart day by day.

I would also like to thank my supervisory committee members, Dr. Richard Morton, Dr. An-Chang Shi, for helpful advice, criticism and support. They are such serious and respected scientists and the happy time during the committee meetings, beer bars and offices are not forgotten. For example, just a few personal chats with Dr. Morton have made a noticeable improvement on my scientific presentation and writing. The chats with Dr. Shi made me know Canadian life and academia as soon as possible, which cannot be taught from books. Both of them also helped me to do research in a professional way, which will have a great impact on my future faculty career in science.

Special thanks to Dr. Herb Schellhorn for his supervision and all kinds of help. The chats on his Rpos data are invaluable for expanding my understanding of bacteria and I hope someday we can work on some question together. His openness on science and life is also impressive and will have a great effect on my future.

I'm grateful for the current and past group members, Nicholas Waglechner, Wenli Jia, Wu Meng for their valuable discussion and help on biology, computer skills and life, and happy moments of lunch time, and hard time in our brightest office till the sunset and midnight; Great thanks to smart Aaron, Osama, Eric and Charmaine for useful discussions.

Many thanks to scientists: Paul Sharp for supplying the codon usage data in 80 bacterial species, Eduardo Rocha for supplying information on bacterial growth times and tRNA gene copy numbers, Hiroshi Akashi for helpful comments on codon usage data in *S. cerevisiae,* Joshua Plotkin for the idea of looking at the effective population size, Barry Cooperman for sharing his understanding on ribosome, and Yingfu Li, Kari Dalnoki-Veress, Brian Golding, Jonathan Dushoff for helpful advice and criticism. Thanks to Dear Dr Donald Sprung for pleasant chats from time to time.

Thanks to Hua Wu for technical support through all these years.

Thanks to Liz, Cheryl, Mara, Tina and Rose for their help on all kinds of documents, letters and reimbursements. They are the best secretaries!

Thanks to all the other people in McMaster communities of Physics, Biology, Biochem and Mathbio. Thanks to all my friends for all the great times we had together.

Finally I would like to thank my family for their eternal support and love. My grandmother, my father and mother, my brothers and my sisters continuously encouraged me to have fun in science. I cannot make any progress without their love and happiness. All my achievements are original from them. Specially thanks to my talented nephews and niece: Jiadong, Lilin, Xin and Zhusong. Their talent in science and art encourage me to work harder and the chats with them make me very happy after a long day work. Their questions those I need to look up in the books or think quite a while always bring me to some new interesting topic about science or life. This thesis is dedicated to all of them.

TABLE OF CONTENTS

Descriptive Note ii

Abstract iii

Acknowledgements v

Table of Contents viii

List of Figures xii

List of Tables xv

Chapter 1 Introduction **1**

1.1 Codon usage bias: mutational and selectional effects . . 1

1.2 Wobble rules for codon-anticodon interactions . . . 5

1.3 Relationship of codon usage bias to expression level . . 6

1.4 Relationship of codon usage to tRNA content . . . 7

1.5 Variation of codon usage bias among organisms . . . 9

1.6 Experiments on the measurement of tRNA concentration . . 9

1.7 Experiments on the measurement of translation rates . . 10

1.8 tRNA modification 13

1.9 Methods to measure codon usage 15

1.9.1 CAI 15

1.9.2 Effective number of codons 16

1.9.3 Correspondence analysis 17

1.9.4 Comparison of high- and low-expression genes . . 18

1.9.5 Comparison of the above methods for CUB . . . 18

1.10 Overview of the thesis 19

Chapter 2 Population genetics and translation kinetics model 22

2. 1 Population genetics: selection-mutation-drift theory . . 22

2. 2 Translational kinetics 27

Chapter 3 Theory of coevolution of codon usage and tRNA genes 32

3. 1 Introduction 32

3.2 Examples showing the dependency of codon bias on tRNA gene copy number 35

3. 3 Coevolution with fixed total tRNA copy number . . . 39

3. 4 Coevolution with variable total tRNA copy number . . . 45

3. 5 Comparison of theory with bacterial codon usage data . . 53

3. 6 Variation of selection strength among organisms . . . 63

3. 7 Discussion 67

Chapter 4 Anticodon-Codon Interaction and tRNA Modification 71

4. 1 Introduction 71

4. 2 Analysis of codon usage in U+C codon families . . . 72

4. 3 Analysis of codon usage in A+G codon families . . . 73

4. 4 Analysis of codon usage in four-codon families . . . 79

4. 5 The role of modified bases at the wobble position . . . 85

4. 6 Conclusion 95

Chapter 5 Translational accuracy 103

5. 1 Introduction 103

5. 2 Sequence data 104

5. 3 Definition of conserved and variable sites 106

5. 4 Simple χ^2 test for codon bias in conserved versus variable site . 108

5. 5 Simple χ^2 test of nonsense error 109

5. 6 A method to test the effect of selection in all codon families . 110

5. 7 Comparison of strength of selection δ for different hypotheses . 112

Chapter 6 Discussion, Conclusion and Future work . . . 118

6. 1 Discussion and Conclusion 108

6. 2 Future work 124

6.2.1 Speed and accuracy variation in different domains . . 124

6.2.2 A kinetic model of translation process . . . 124

6.2.3 MCMC simulation on anticodon-codon interaction . . 125

References 127

Appendix List of species used in the thesis **136**

LIST OF FIGURES

Chapter 1

Fig. 1.1 Standard genetic code table 2

Fig. 1.2 Translation kinetic model 6

Fig. 1.3 Translation process 12

Fig. 1.4 tRNA modification 14

Chapter 3

Fig. 3.1 (a) Expected frequency ϕ as a function of K
(b) Mean translation time per codon as a function of ϕ for each tRNA gene combination 42

Fig. 3.2 Frequency of the C codon in U+C families as a function of K for varying numbers N_G of tRNA copies . . . 47

Fig. 3.3 Stable solutions for codon frequency as a function of K in A+G families 50

Fig. 3.4 As Figure 3.3 except that $\theta = 0.7$ 51

Fig. 3.5 Variation of codon usage ϕ with the GC content of the mutation process θ 52

Fig. 3.6 (a) Relationship between the minimum doubling time of bacteria and the estimated value of the translational cost parameter
(b) Dependence of the total number of tRNA genes in bacterial

rate 115

Fig 5. 4 No correlation is found for δ_C, δ_V and growth rate. The values of δ_C and δ_V and the difference between δ_C and δ_V are small, which means only a weak selection exists in the conserved sites . 116

Fig 5. 5 No correlation is found between growth rate and the δ values of the 3' half and 5' half of genes in bacteria. The values of δ_C and δ_V and the difference between $\delta_{3'\,half}$ and $\delta_{5'\,half}$ are small, which means only a weak selection exists in the two halves. In other words, non-sense error is weak 117

LIST OF TABLES

Chapter 1

Table 1.1 tRNA molecules, tRNA gene copies of Bacteria . . 8

Chapter 3

Table 3.1 Examples of codon usage, estimated selection strength, and tRNA gene copy numbers in representative species for two amino acids with U+C codon families. 36

Table 3.2 Examples of codon usage, estimated selection strength, and tRNA gene copy numbers in representative species for two amino acids with A+G codon families. 37

Table 3.3 Comparison of theory for U+C codon families with observations in bacteria 58

Table 3.4 Comparison of theory for A+G codon families with observations in bacteria 58

Chapter 4

Table 4.1 Codon usage and tRNA content in A+G families . . 97

Table 4.2 Codon usage in four-codon families where only wobble-U tRNAs are present 98

Table 4.3 Codon usage in four-codon families with both wobble-U and wobble-G tRNAs 99

Table 4.4 Codon uses in four codon families with combinations of wobble-C tRNAs with wobble-U and wobble-G tRNAs . . 100

Table 4.5. Codon usage in four codon families involving wobble-A (or I) tRNAs 101

Table 4.6. Summary of the modifications of different base . . 102

Chapter 5

Table 5.1 χ^2 test on all three sets of genes 109

Table 5.2 Comparison of details of χ^2 test for conserved and variable sites with a percentage of 80 and for high and low expression genes in all pecies 109

Chapter 1

Introduction

This chapter is a review of the genetic code table, codon usage bias, translational selection and tRNA modification. An outline of the thesis is also given at the end.

1.1 Codon usage bias: mutational and selectional effects

One of the greatest discoveries in the past one hundred years is the elucidation of DNA structure by Watson and Crick (1953), which is the beginning of an age of molecular biology. Crick proved that three DNA bases are needed to form a codon for the decoding of each of the 20 amino acids and a signature for termination of translation (Crick *et al.*, 1961). There are 4 different bases in DNA structures. So in total we have 4^3=64 possible codons, which are displayed in the genetic codon table (Fig 1.1). There are many ways to draw codon tables for different uses and the table here is the standard one. Codons are degenerate, *i.e.* usually more than one codon is used to encode each amino acid. Codons encoding the same amino acid are called synonymous codons. The frequencies of these synonymous codons, or codon usages, are generally not equal. This phenomenon is called Codon Usage Bias (CUB). Codon usage differs among codon families for different amino acids in a same organism and as well as among organisms.

Codon usage and CUB have been studied for a long time. Two years after Crick's experiment, Ames and Hartmann first pointed out that synonymous codons should have an effect on the regulation of gene expression (Ames and Hartmann, 1963). Scientists who believed the neutral theory of evolution thought that the synonymous codon usages

were affected just by mutation pressure (King and Jukes, 1969). However, Clarke (1970) argued that CUB could also be due to selection pressure. Now it is well accepted that CUB should be due to both effects: the mutation effect that leads to the variation of GC composition and a selection effect arising from optimization of translation speed and accuracy.

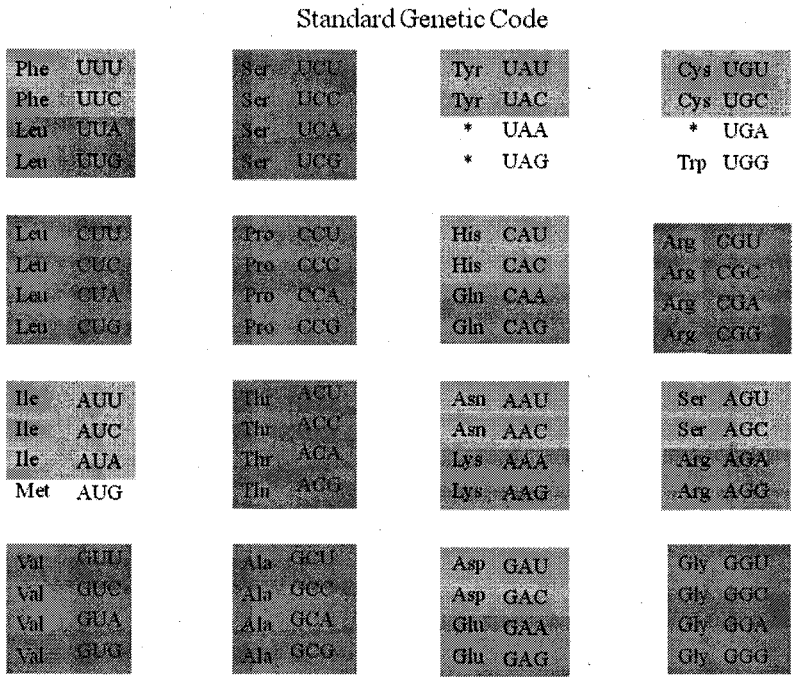


Fig. 1.1 Standard genetic code table where codons are grouped into: U+C codon family (also called as pyrimidine families), A+G codon family (purine famililes) and 4-codon families.

The GC composition can be quite different for different organisms due to mutation bias. Knight *et al.* (2001) studied the GC composition in a large set of genomes of 311 bacteria, 28 archaea and 257 eukaryotes. They found there are strong correlations between GC content and codon usage, and between GC content and amino-acid usage.

Another good example of mutation bias is shown in *Mycoplasma genitalium* (Kerr *et al.* 1997), whose GC content varies randomly with genome locations.

The speed hypothesis assumes that the rate of protein synthesis is a significant factor limiting the growth rate of cells. Fast translation should be an advantage for bacterial growth because it will allow more rapid protein synthesis by a limited number of ribosomes. According to this hypothesis, codons that are more rapidly translated are used more often in an organism in order to reduce the time and effort spent on translation (Sharp *et al.* 1988, 2005; Akashi, 2003; Dos Reis *et al.* 2003). In many organisms, frequent codons are the ones that match the tRNAs that are most abundant in the cell (Ikemura, 1981, 1985; Percudani *et al.* 1997; Duret, 2000) because generally a higher tRNA concentration leads to a more rapid translation. In cases where there is only one type of tRNA for an amino acid, the more frequent codon is usually the one that is more rapidly translated by that tRNA.

The accuracy hypothesis assumes that accurate translation should be an advantage because mistranslation will lead to an unfinished amino acid chain (nonsense error) or a wrong amino acid (missense error), which is a waste of energy and sometimes even toxic to the cell. According to this hypothesis, the more frequent codons should be those that are more accurate translated (Eyre-Walker 1996; Gilchrist *et al.*, 2009). A nonsense error is a premature termination of translation before ribosome reaches the end of the mRNA. If selection acts against nonsense errors, more accurate codons should be more frequent in the end half (3' half) of any sequence than in the beginning half (5' half).

Sites in a protein where an amino acid is evolutionarily conserved over a range of species are presumably those that are important for protein folding and function. Missense errors at conserved sites should therefore be a disadvantage. Thus, if selection against missense errors is an important cause of CUB, we expect accurate codons to be more frequently at conserved sites (Akashi, 2003; Eyre-Walker 1996; Eyre-Walker 2006).

Proteins such as ribosomal proteins and elongation factors are present in high concentrations in the cell. The sequences of these highly expressed genes are found to have strong codon usage bias with respect to the majority of low expression genes (Sharp and Li, 1986). This is to be expected if selection acts on speed because the time saved by using a faster codon is proportional to the number of times the gene is translated. It is also to be expected if selection acts on accuracy, because the number of mistranslated proteins produced (Drummond and Wilke, 2008) will also be proportional to the number of times the gene is translated.

Grantham *et al.* (1980; 1981) studied codon usage and pointed out that codon usage is an organism-specific character and genes of any organism can be grouped based on this measure. Ikemura *et al.* (1981; 1985) observed that codon usage was correlated with tRNA abundance and gene expression levels. After more researches on the model organisms such as *Escherichia coli* and *Saccharomyces cerevisiae* (for example, Sharp and Li, 1986, 1987), Bulmer (1991) summarized the main features of CUB as the following: 1). tRNA abundance is positively correlated with tRNA copy number. 2) tRNA abundance is positively correlated with the usage of corresponding codon in genomes. 3). For codons interacting with the same tRNA, a codon is usually preferred if

the 3rd base of the codon and the 1st base of the anticodon form a Watson-Crick pair. 4). Usually CUB happens in high expression genes.

Up to now quite a few experiments have been done in model organisms which support either the translation speed hypothesis or the accuracy hypothesis as the cause of CUB apart from mutation pressure. However, there is no systematic analysis of the origin of CUB and no qualitative or quantitative ways to interpret CUB in a large number of genomes across all bacteria, taking both effects of speed and accuracy into account. This is the motivation and main content of my thesis. Ikemura's observation (1981) on the correlation of preferred codon and Watson-Crick pair is another interesting point, which will be discussed in Chapter 4.

1.2 Wobble rules for codon-anticodon interactions

During translation, a tRNA molecule charged with an amino acid goes into ribosome and an amino acid chain is synthesized using mRNA as a template. Generally a tRNA molecule can translate more than one codon because of the wobble rules (Crick, 1966). The first and second positions of a codon always follow the Watson-Crick pairs (A-U, U-A, G-C and C-G pairs) during translation. But the third position is flexible. The first base in the anticodon is position 34 in the standard tRNA alignment (Sprinzl and Vassilenko 2005), and this is also called the wobble position. The base on the wobble position of a tRNA is flexible and able to pair with a base on the third position of any synonymous codon. These codon-anticodon pairs have been reviewed by Grosjean *et al.* (2010). All observed pairs are discussed in that review. In Fig 1.2 I summarize them in our translation

model. This universal translation model is based on the fact that a codon can be translated by different tRNA molecule and a tRNA molecule can translate different codons in the genetic codon table, which is unique and not taken into account in other models.

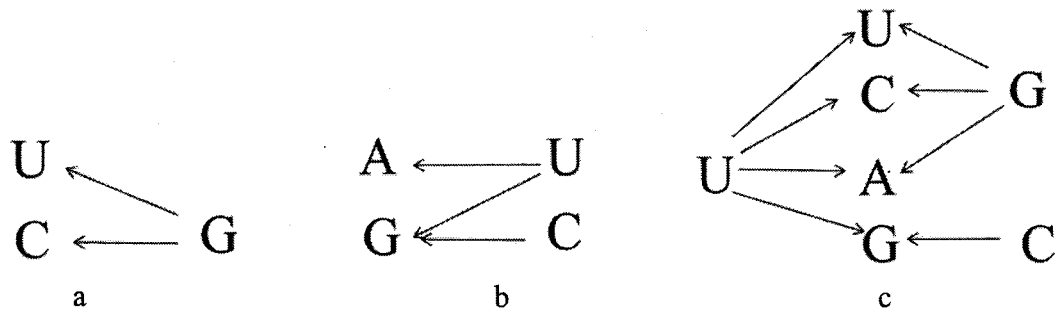


Fig. 1.2 Translation kinetic model shows pairing possibilities for three types of codon families: (a) U+C families; (b) A+G families; (c) four-codon families.

In this thesis, I group codon families according to the base at the third position. For U+C codon families, only a G-tRNA is needed for translation of both codons. For A+G codon families, a U-tRNA is enough for translation but sometimes an extra C-tRNA is also present. The four codon families are more complex since a U-tRNA can fulfill the translation alone or with the help of either a G-tRNA or C-tRNA or both. So all possible anticodon-codon interactions are presented in the model shown in Fig 1.2 and I will use it to construct our theory for selection on translational speed in Chapter 2.

1.3 Relationship of codon usage bias to expression level

Gene expression is a process where the genetic information (stored within DNA) is used for the production of proteins. Gene expression level is correlated to CUB (Sharp and Li, 1987; Henry and Sharp, 2007). In some species, expression levels of all genes have been

measured, either at mRNA level or protein level. However, systematic measurement has not been done for large numbers of genomes. Ribosomal proteins can be considered as high expression genes since the mass of ribosome proteins is around 25% of the total mass of a bacteria cell (Arnold and Reilly, 1999) and the expression level of these genes is always high. Akashi (2003) studied the codon usage in *Saccharomyces cerevisiae* with tRNA gene numbers, amino acid compositions and DNA microarray expression data collected from many publications. A strong correlation between the expression level and CUB was found. In other words, CUB was much higher in high expression genes than that in low expression genes. He argued that a strong selection acted on these high expression genes to make translation process fast and accurate and CUB is a result of this selection.

1.4 Relationship of codon usage to tRNA content

The total number of tRNA genes and the number of copies of each type of tRNA gene in the genome vary in different organisms. The smallest complete set of tRNAs occurs in animal mitochondria, where there are only 22 tRNA molecules, each with a single copy. In bacteria there are always more than this, as is shown by the examples in Table 1.1 (Full details of the numbers of tRNAs in all the bacteria analyzed here are shown in the Appendix at the end of this thesis), where both the tRNA gene copy number and tRNA anticodons vary a lot. The model organism *Escherichia coli* has 86 tRNA genes. More tRNA genes are needed for eukaryotes: There are 275 tRNA genes in the genome of *Saccharomyces cerevisiae*, while for *Caenorhabditis elegans*, 659 tRNA genes. All these

data can be found at http://lowelab.ucsc.edu/GtRNAdb/, where many eukarya and bacteria species are given. Rocha (2004) studied the tRNA gene copy number of 102

Organism	#gene copies	#different anticodons
Buchneria aphidicola	31	29
Agrobacterium tumefaciens	53	39
Corynebacterium glutamicum	60	42
Escherichia coli	86	39
Vibrio vulnificus	112	32

Table 1.1 tRNAs, tRNA gene copies of Bacteria. All data are from website http://lowelab.ucsc.edu/GtRNAdb/

bacterial species with their growth rates (see Appendix). Growth rate is a measurement to show how fast bacteria grow under a certain condition and the value is the inverse of minimum generation time or doubling time. He found that for most bacteria and small eukaryotic genomes the number of tRNA anticodons decreases but the number of tRNA gene copies increases with the minimal generation times. The CUB is also stronger in organism with shorter generation time. The minimum doubling time or generation time he used was collected from publications and his personal communication, as he stated in his

paper. It was measured in optimal growth conditions such as rich resource and optimized temperature. For example, the optimized temperature of *Escherichia coli* is *37°* C. Generally, the media for different organisms were different. If an optimized doubling time was given by different groups, the smaller value was used. Codon usage may also be related to genome size.

1.5 Variation of codon usage bias among organisms

As the sequence data were available for many organisms, Sharp *et al.* (2005) explored CUB across a wide range of bacteria. They randomly selected 80 bacteria species from all bacterial subgroups and calculated their CUB, which vary a lot among species. They reported that 70% of the genomes show selection to be effective but no significant selection exists in the rest genomes in their study. They also found the values of their selection parameter are highly positively correlated with both the number of rRNA operons and the number of tRNA genes. They interpreted this as a sign of selection acting on speed. Rocha (2004) also studied the variation of CUB and found that it was positively correlated to the growth rate of species. The potential explanation is that high expression genes are supposed to be used to produce large amount of proteins in a genome and mainly determine the growth rate of an organism. For fast-multiplying bacteria, translation time is a limiting factor and short translation time is likely an advantage. So strong bias should exist in fast growing bacteria and eukaryotes as an effect of selection.

1.6 Experiments on the measurement of tRNA concentration

Two-dimensional polyacrylamide gel electrophoresis is a widespread method to separate tRNAs and measure their concentrations. Ikemura (1981) used it to separate 26 tRNAs of *Escherichia coli* and obtain their relative abundance. He found a strong correlation between tRNA abundance and codon usage, especially in ribosomal proteins. He concluded that gene number was a major factor to determine the contents of tRNA. Ikemura (1985) reviewed codon usage and tRNA content in *Escherichia coli* and yeast, where he confirmed his former results and suggested the cause was optimization of protein production. He also concluded that unicellular eukaryote had similar behaviors in codon usage bias as prokaryotes, but multicellular eukaryotes appeared to be less influenced by translational selection. In *Caenorhabditis elegans* tRNA gene number and codon usage are also found to be co-adapted (Duret, 2000). Now it is widely accepted that tRNA gene number, or tRNA concentration and codon usage are co-adapted for optimization of translation or protein production.

Dong *et al.* (1996) used two-dimensional polyacrylamide gel electrophoresis to measure the abundance of almost all tRNA species in *Escherichia coli* at different growth rates, which were varying from 0.4 to 2.5 doublings per hour. This experiment clearly showed that there was a positive correlation between tRNA concentration and tRNA gene copy number. Similar conclusions were obtained for yeast (Percudani *et al.* 1997), which may imply that the correlation is universal in unicellular organisms. Ardell and Kirsebom (2005) investigated the effect on expression of transcription of tRNAs in operons of several genes. They reported that geomic location and strandedness of tDNA operons could also have an effect on the tRNA expression.

1.7 Experiments on the measurement of translation rates

In this thesis I will argue that selection on translational speed is an important cause of CUB. Therefore in this section I will discuss experiments that study the kinetics of translation, which is a topic that has been studied for more than 30 years (Ninio, 2006). The translation process is described in Fig. 1.3. The initial binding happens as tRNAs enter ribosome in the form of a ternary complex, consisting of the aminoacylated tRNA, an elongation factor protein EF-Tu, and a GTP molecule. Then the codon-anticodon interaction induces a conformational change in EF-Tu, which triggers GTP hydrolysis. This is called GTPase activation (GTPase is defined as the hydrolase enzyme that binds and hydrolyzes GTP). GTP hydrolysis and accommodation of aa-tRNA then happen at P-site, where an amino acid peptide bond is formed and the new amino acid is joined with the existing amino acid chain. The tRNA then exits from the E-site. Although the general picture is well accepted, different versions of kinetic details have been presented by different groups. Here two of them are described. The first one is given by Rodnina *et al.* (2001). They built a model of kinetic pathway of A-site binding and measured the rate constant of each step. Their results showed that two key phases related to translational selection were GTPase activation and GTP hydrolysis, and accommodation of aa-tRNA and formation of peptide bonds. The interesting point was that the rate constant of initial binding and codon recognition showed little difference for cognate and near-cognate groups, which led to an unexpected conclusion that there was no selection at this stage. In their experiments, a cognate codon was the codon that matched a tRNA exactly (this

tRNA was called the cognate tRNA for that codon) and a near-cognate codon was the codon whose 1st or 3rd base did not match the tRNA.

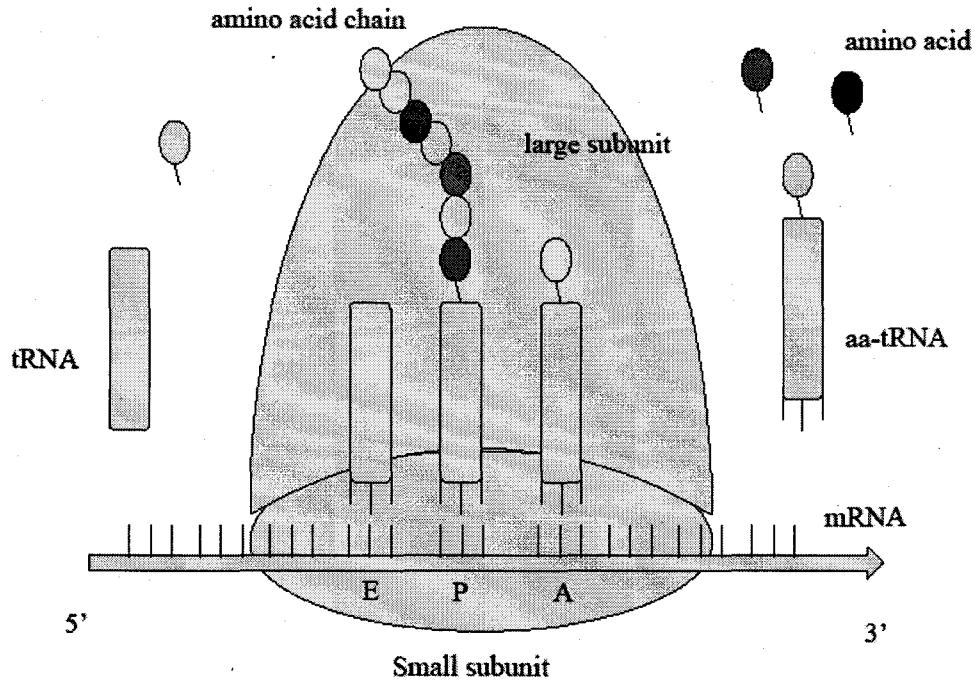


Fig 1.3 Translation process. tRNAs make amino acid chain using mRNA as a template.

However, a later experiment doesn't agree with that of Rodnina's group (Blanchard *et al.*, 2004). They used single-molecular fluorescence intensity to measure the translation rates for different steps in translation. They found that within the small subunit decoding site in ribosome, the tRNA anticodon and mRNA codon contacted each other with a proper conformation in initial binding process. So a proper orientation of EF-Tu-GTP-aa-tRNA and P-site tRNA was fixed first. They also believed the cause of this

initial selection was supposed to be related to the structure conformation and folding reaction between EF-Tu-GTP-aa-tRNA and ribosome as well. This was a codon recognition step not found by Rodnina.

Ninio (2006) reviewed experiments those measured translation rates and schemes of kinetics of tRNA selection. He pointed out the following facts those made a unique scheme hard to be obtained from today's experiments. Some step between accommodation of aa-tRNA and formation of peptide bond existed in nature, which could not be detected due to the limit of today's lab technique. There was a big difference in the published data within and between lab groups and the data themselves were not clear to support a unique scheme. For example, Rodnina's group had two sets of data on the GTP hydrolysis rates and neither of them agreed with those of Blanchard's group. This was partly due to the fact that lab groups had a predetermined scheme before their measurements.

1.8 tRNA modification

There are many cases where the wobble base of tRNA is modified in a way that influences anticodon-codon pairing (Curran, 1998; Agris 2004, 2008). tRNA modification is important because it has huge effect on anticodon-codon interaction. Although a considerable amount is known about the functions of these modifications in some species, there is no systematic information about modifications for all the genomes analyzed in this thesis. The modifications (see figure 4.2 for the chemical structures and Fig. 1.4 for

the positions of modifications) and their effects will be discussed in Chapter 4. Here I just summarize briefly the most frequent modifications.

In U+C families, the G base at the wobble position is modified to queuosine, Q, in tRNAs for Tyr, His, Asn and Asp in all domains of life (Romier *et al.* 1998). In tRNAs for four-codon families, the U base at the wobble position is modified to 5-methoxyuridine (mo^5U) or uridine-5-oxyacetic acid (cmo^5U) in most bacteria (Curran, 1998; Agris, 2004, 2008; Jühling *et al.*, 2009). I denote this class of mutations as xo^5U. In tRNAs for two-codon families, the U base usually has a 5-methylaminomethyl modification and often a 2-thio modification as well. I denote this class as xm^5U. The two classes of modified U bases function in different ways, which will be discussed in Chapter 3. Another case in four codon families is that A is almost always modified to inosine (I) in mature tRNA, which will also be discussed.

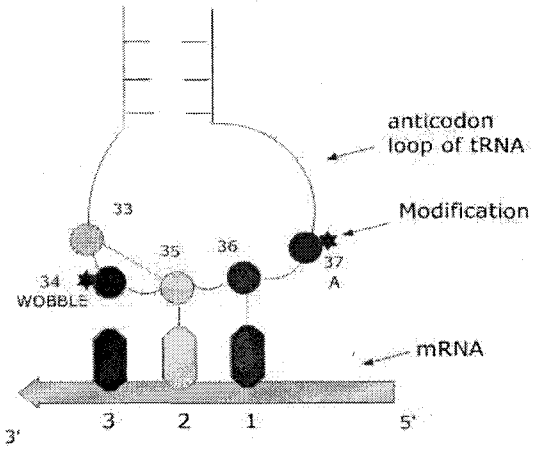


Fig 1.4 tRNA modifications.

I have focused on modifications at the wobble position (tRNA position 34) because these have a direct interaction with the third codon position. However, base modifications in other positions are also significant in terms of translation and possibly codon usage. In particular, modifications often occur at position 37 (the base that follows the anticodon, as shown in Fig 1.4). Removal of these modifications has been shown to have a detrimental effect on either speed or accuracy of translation in some cases (Yarian, 2002; Agris 2004, 2008). Base modifications at positions 34 and 37 have also been found to be important for proper translocation of a tRNA from the A site to the P site in the ribosome (Phelps *et al.* 2004), and in maintenance of correct reading frame (Urbonavicius *et al.*, 2001).

1.9 Methods to measure codon usage

1.9.1 CAI

Sharp and Li (1987) developed codon adaptation index (CAI) to describe CUB. I will reproduce their definition here. Let ϕ_i be the frequency of codon i in the set of high expression genes used as a reference. Let ϕ_i^{max} be the frequency of the highest-frequency codon for the same amino acid. A weighting factor $w_i = \phi_i/\phi_i^{max}$ is then defined, such that $w_i = 1$ if i is the most frequent codon for that amino acid, and $w_i < 1$ if i is a less frequent codon. For a gene of length L codons containing n_i copies of codon i, the CAI is defined as the geometric mean of the weights of codons in that gene.

$$CAI = (\prod_i w_{ij}^{n_i})^{\frac{1}{L}} = \exp(\frac{1}{L}\sum_i n_i \ln w_i) \qquad (1.1)$$

This index is the most widely used measurement of CUB ever since they used it in the high and low expression genes in *Escherichia coli* and yeast in 1987.

1.9.2 Effective number of codons

Bennetzen and Hall (1982) observed that for some organisms, such as yeast, only 22 codons from the standard 61 codons were used to do the translation for high expression genes, while all codons were used for low expression genes. This means the codons are used non-evenly in these two groups of genes. Wright (1990) studied the usage of codons which were used to translate amino acids in all kinds of organisms. He developed the effective number of codons used in a gene, N_C, as a measure of CUB.

Let there be n_i copies of codon i in a gene sequence and let n be the total number of copies of codon coding for an amino acid. If two of these codons are chosen at random, the probability that they are the same is

$$F = \sum_i \frac{n_i(n_i-1)}{n(n-1)} \qquad (1.2)$$

If only one codon is used for an amino acid, $F = 1$. If n is large and all codons for this amino acid are used with equal frequency, then $F = 1/q$, where q is the number of codons for this amino acid. For a partially biased set of codon F varies between 1 and $1/q$. So we can regard 1/q as the effective number of codon that is used for an amino acid.

Wright defined N_C as the effective number of codons used for all amino acids in the following way. There are five types of amino acids in a tandard code: 2 amino acids have just one codon, 9 for two, 1 for three, 5 for four, and 3 for six. For each type of amino acid, Let F_1,F_2, F_3, F_4 and F_6 be the mean values of F for amino acids with q= *1,* 2,

3, 4 and 6 (F_1=1 is the easiest case). Then N_C is obtained by summing the effective number of synonymous codons used by each of 20 amino acids.

$$N_C = 2 + \frac{9}{F_2} + \frac{1}{F_3} + \frac{5}{F_4} + \frac{3}{F_6} \qquad (1.3)$$

Theoretically Nc will equal to 61 if all codons are uniformed used, while 20, for a completely biased codon usage. An effective number of codons gives a direct measure of the extent of codon usage in a gene. Wright used Nc to study the CUB in organisms such as *Homo sapiens, Saccharomyces cerevisiae, Escheriehia coil, Bacillus subtilis, Dictyostelium discoideum, and Drosophila melanogaster.* He concluded that Nc was a very good method for CUB because it was easy to show the even or non-even codon usage in any organisms.

1.9.3 Correspondence analysis

Based on principal component analysis (Karl Pearson, 1901), Jean-Paul Benzécri (1973) developed correspondence analysis, which is a technique to study the correspondence between the rows and columns of a data set in a given table. In biology it is often used to separate genes in different groups and study the properties or correlations of these gene groups. Dos Reis *et al.* (2003) compared data from different researchers (Selinger, 2000; Bernstein, 2002; Glasner, 2003) and used correspondence analysis on these data. All the data were separated into three groups: high expression genes, moderate expression genes and AT-rich genes with low CUB. They were surprised to find that the AT-rich genes with low CAI were highly expressed. This was not easy to understand. They finally turned to the assumption that all AT-rich genes were from horizontal transfer. Since these

high expression genes were foreign genes, they had low codon adaptation index and tended to change their GC content to that of host (dos Reis, 2003)

1.9.4 Comparison of high- and low-expression genes

Sharp *et al.* (2005) estimated the selection strength on codon usage for 80 bacterial species using population genetic theory. An organism-specific parameter for selection was also developed and the selection was estimated from genetic information from both low and high expression genes. The selection strength of each amino acid was calculated and selection strength across amino acids was obtained for different species by taking into account the weight of codon numbers of each amino acid in high expression genes. This is the method that will be used in my thesis, where it will be expanded and can be used for not only U+C codon families but also all codon families. A detailed theory can be found in Chapter 2.

1.9.5 Comparison of the above methods for CUB

The CAI is the first and most widely used measure of CUB that accounts for selection. It is a direct index of CUB in any group of genes using high expression genes as a reference. Generally it is used to compare CUB in different groups of genes in one organism. It is also widely used as an indication of expression level when the measurement of expression level is not available. However, it cannot be used to compare selection among different genomes because the codon frequencies in high expression genes are different for different organisms and hence the w_i values in equation 1.1 are different. The effective number of codons is a way to describe even or non-even usage of all codons without

consideration of selection or mutation. Compared with the above two methods, the one used by Sharp and discussed in Chapter 2 is better because: 1) It can be used to compare selection strengths between different organisms. 2) The selection strength parameter comes from the information of both low and high expression genes. 3) It considers the effect of selection. 4) It distinguishes the preferred codon from the frequent codon: A frequent codon is a codon used most either in high or low expression genes, while a preferred codon is the one whose codon usage increases from low expression genes to high expression genes.

1.10 Overview of the thesis

Chapter 2 introduces the population genetics and translation kinetic model that are used throughout the whole thesis.

Chapter 3 describes the theory of coevolution of codon usage and tRNA genes. The theory explains the observation that CUB varies in different amino acids even in the same organism. Based on the theory of selection – mutation – drift and the translation model, a theory of multiple stable states (MSS) is built to explain the phenomenon that in A+G codon families preferred codons can be different for different amino acids in the same genome. It reveals that both codon usage and tRNA copy number have to co-adapt to each other to make the translation process fast, which causes the preferred codon varies in different codon boxes. In addition to S, a new selection parameter K is introduced as a better measure of organism-specific selection since it excludes the effects of tRNA concentration and anticodon-codon interaction, which makes the comparison of selection

strength between different species more sensible. The calculation of K is performed in six U+C families and K values are plotted against the minimum doubling time T of 80 species. The linear regression shows that there is a strong correlation between K and $1/T$ with a large statistic significance which supports the hypothesis that translation speed is the main reason of codon usage bias and K is a better parameter as a measurement of selection strength. The effective population size N_e has also been studied to some extent and the conclusion is that K is independent of $N_e u$. Most of the contents of this chapter are from Higgs and Ran (2008).

In Chapter 4 the selection strengths of all codon families are calculated and the analysis of these data reveals the relative magnitude of anticodon-codon interactions, which are explained by tRNA modifications. The codon usage data in each codon family of all species are studied without either speed or accuracy hypothesis. The selection strength S, which is a measurement of codon bias by counting the numbers of codons in high and low expression genes, is calculated following the method introduced by Sharp *et al.* (2005). Then data are grouped according to different combinations of tRNAs and the averages of codon usage for each combination are taken. These data are joined with the translation model and the predictions about anticodon-codon interaction are made that can be tested by experiments. The interesting phenomenon is that anticodon-codon interactions are different between two codon families (the U+C and A+G codon families) and four-codon families. Differences are due to tRNA modifications, mainly on position 34 of tRNA anticodon loop. Different modifications are introduced for different codon families which follow different translation patterns during evolution, which help

translation fast and accurate. This is shown in Fig. 4.2 and is the central picture of this chapter. Most of the contents of this chapter are from Ran and Higgs (2010).

Chapter 5 presents the results of sequence analysis of 47 ribosomal protein genes in bacteria. The effects of translation accuracy and speed on CUB are studied. In Chapter 3 and 4 only the speed hypothesis is used to explain the codon usage data, which can be also due to other possible reasons, including strand bias, mutation bias and selection on accuracy. However, the data fit the theories based on speed hypothesis quite well. As a further investigation in this chapter, a comparison between the effects of translation speed and accuracy is made, which makes it possible to tell which contributes more to codon usage bias. This is done by two tests: the comparison of CUB in the conserved site versus the variable ones on the amino acid chain in high expression genes; and the comparison of CUB in the 5' and 3' halves of genes. The conclusion is that there is a weak additional selection for accuracy acting in combination with selection for speed. Statistical analysis shows that the C codon is selected on the grounds of both speed and accuracy in the U+C codon families. In other words, the fast codon is also the accurate one, which is possibly a general rule.

Chapter 2

Population Genetics and Translation Kinetics Model

Preface

A population genetics model of codon usage bias in two codon families is presented and expanded to work on both two and four codon families. A universal translation kinetics model is also introduced. These will be used in the following chapters.

2. 1 Population genetics: selection-mutation-drift theory

Li (1987), Shields (1990) and Bulmer (1991) built a model showing how codon usage depends on selection, mutation, and drift (SMD). Consider a site at the third position in a codon belonging to a U+C two-codon family, such as that for Phe, for example. Let π be the GC content of the genome that would arise from mutation alone, *i.e.* let the rate of mutation from U to C be $u\pi$ and the rate of the reverse mutation be $u(1-\pi)$. Here u determines the absolute values of mutation rates. In the absence of selection, the relative frequencies of U:C will be $(1-\pi):\pi$. In the presence of selection, we will define the fitness of the U codon as $w = 1$, and let the fitness of the C codon be $w = 1+s$. The magnitude of genetic drift is controlled by the effective population size, N_e. The theory considers the limit where $N_e u$ is small, so that for most of the time, the population is dominated by one nucleotide or the other, but it occasionally makes transitions between the two.

In a population where almost all individuals have a U at a particular third position site, the net rate of creation of C mutations is $u\pi \times N_e$. The probability that any of these mutations becomes fixed in the population is (Kimura, 1962)

$$P_{fix} = \frac{1-\exp(-2s)}{1-\exp(-2N_e s)}. \tag{2.1}$$

The net rate of transition of the population from the U to the C state is

$$R_{UC} = u\pi N_e P_{fix} = u\pi F(S), \tag{2.2}$$

where

$$F(S) = \frac{S}{1-\exp(-S)}. \qquad S = 2N_e s \tag{2.3}$$

Here, it has been assumed that selection on synonymous sites is weak ($s << 1$), so that the numerator of equation 2.1 is approximately $2s$. In this case, the effectiveness of selection is controlled by the parameter $S = 2N_e s$, and the parameters s and N_e do not appear separately in the equation. The net rate of transition from the C to the U state is $R_{CU} = u(1-\pi)F(-S)$. Let the relative frequencies of C in the presence of selection is $\phi(S)$; then that of U is $1-\phi(S)$. At equilibrium under SMD we have

$$\frac{\phi(S)}{1-\phi(S)} = \frac{R_{UC}}{R_{CU}} = \frac{u\pi F(S)}{u(1-\pi)F(-S)} = \frac{\pi(\exp(S)-1)}{(1-\pi)(1-\exp(-S))} = \frac{\pi\exp(S)}{1-\pi}. \tag{2.4}$$

Note that the dependence on S reduces to a very simple exponential function in the last step above. Rearranging this gives:

$$\phi(S) = \frac{\pi\exp(S)}{\pi\exp(S)+1-\pi}. \tag{2.5}$$

In the U+C families, the C codon is almost always preferred, *i.e.* $S > 0$, and $\phi(S) > \pi$. This theory is explained in Sharp's paper (2005) and we just summarized it here.

The above formula is used to calculate the selection strength directly from DNA sequences. In the following two chapters, we assume that mutation just acts on low expression genes while both selection and mutation should exist in high expression genes and the selection is mainly due to speed. Then selection is significant principally on a relatively small number of genes, which are much more highly expressed than the average. Genes such as ribosomal proteins and elongation factors are presumed to be highly expressed in all organisms, and codon frequencies in these genes should be indicative of those in genes in which selection is operating. On the other hand, the levels of expression of the majority of genes in the genome are much lower, and the strength of selection on the majority of genes may be negligible in comparison to that on the small number of very highly expressed genes. Using these assumptions, Sharp *et al.* (2005) estimated the strength of selection acting in the U+C codon families in bacterial genomes. We will use the same method. Let n_U^{low} and n_C^{low} be the number of U and C codons in a U+C codon family in the whole genome. These are labelled 'low' because the majority of genes are assumed to be expressed at a low level. Let n_U^{high} and n_C^{high} be the numbers of codons in a small set of genes that are presumed to be highly expressed. If selection is negligible on the low expression genes, we expect codon usage to depend only on the mutation parameter π:

$$\frac{n_C^{low}}{n_U^{low}} = \frac{\pi}{1-\pi}, \quad \text{and} \quad \pi = \frac{n_C^{low}}{n_U^{low} + n_C^{low}}. \tag{2.6}$$

In the high expression genes, codon usage depends on θ and S via the $\phi(S)$ function:

$$\frac{n_C^{high}}{n_U^{high}} = \frac{\phi(S)}{1-\phi(S)}, \quad \text{and} \quad \phi(S) = \frac{n_C^{high}}{n_U^{high} + n_C^{high}} . \tag{2.7}$$

Rearranging (2.4) and using (2.6), (2.7), we have

$$S = \ln\left(\frac{\phi(S)}{1-\phi(S)}\frac{1-\pi}{\pi}\right) = \ln\left(\frac{n_C^{high} n_U^{low}}{n_U^{high} n_C^{low}}\right). \tag{2.8}$$

Thus, both π and S can be estimated by simple counting of codons. For the A+G codon families, the equations are equivalent with subscript G replacing C and subscript A replacing U. In this case S is the selective advantage of the G codon with respect to the A. S is positive when G is preferred and negative when A is preferred.

The above are cases for two codon families. Now we will summarize it in a way that is expanded to four-codon families. For each codon i in a given family we set the fitness to be $1+s_i$. We choose a reference codon and define $s_i = 0$ for the reference codon. For the other codons, s_i is the selective advantage or disadvantage of this codon with respect to the reference. Let π_i be the frequency of base i under the mutation process. The mutation rate from the reference codon to codon i is $u\pi_i$ and the reverse mutation rate is $u\pi_{ref}$. The value of u determines the absolute values of the mutation rates, but the expected frequencies of the codons do not depend on u. Let ϕ_i and ϕ_{ref} be the expected frequencies of the two codons under mutation-selection-drift balance. Using the population genetics theory described in the papers cited above, we find

$$\frac{\phi_i}{\phi_{ref}} = \frac{\pi_i \exp(S_i)}{\pi_{ref}}, \tag{2.9}$$

and $$\phi_i = \frac{\pi_i \exp(S_i)}{\sum_j \pi_j \exp(S_j)}, \tag{2.10}$$

where $S_i = 2N_e s_i$., and N_e is the effective population size. If a different codon were chosen as reference, this would shift all the S_i values up or down by a constant, but this would make no difference to ϕ_i. Therefore, it makes no difference which codon is used as reference. Then high expression genes will be identified in which selection on codon usage is presumed to be significant and where codon frequencies should be given by Equation 2.10. The majority of genes in the genome is assumed to be low expression genes where selection is negligible and where codon frequencies are equal to the frequencies under mutation, π_i. Let n_i^{low} and n_i^{high} be the numbers of occurrences of codon i in a codon family for one particular amino acid in one particular genome for the high and low expression genes, respectively. If the low expression genes are only influenced by mutation, then the π parameters can be estimated from the observed codon counts in the low expression genes, *i.e.*

$$\pi_i = \frac{n_i^{low}}{\sum_j n_j^{low}}. \tag{2.11}$$

Similarly, the frequencies under selection can be estimated from the observed codon counts in the high expression genes,

$$\phi_i = \frac{n_i^{high}}{\sum_j n_j^{high}}. \tag{2.12}$$

Now, rearranging (2.9), and using (2.11) and (2.12) gives:

$$S_i = \ln(\frac{\phi_i}{\phi_{ref}}\frac{\pi_{ref}}{\pi_i}) = \ln\left(\frac{n_i^{high} n_{ref}^{low}}{n_{ref}^{high} n_i^{low}}\right). \quad (2.13)$$

Thus, the selection parameters S_i can be estimated directly from the codon counts.

2. 2 Translational kinetics

When considering how to incorporate translational kinetics into a theory for codon usage we need to account for two principal facts. Firstly, preferred codons correspond to tRNAs that are more frequent; thus we assume the rate of translation of a codon should increase with the frequency of its cognate tRNAs. Secondly, selection occurs between different codons that are translated by the same tRNA; so we assume translation rates must depend on the individual anticodon-codon pair. The simple assumption that incorporates these factors is to suppose that the rate at which tRNAs of type i translate codons of type j can be written as $r_{ij} = C_i k_{ij}$, where C_i is the tRNA concentration and k_{ij} is a rate constant specific to the codon-anticodon pair. In the case that more than one tRNA translates the same codon, the rate of translation of codon j is $r_j = \sum_i C_i k_{ij}$, where the sum is over all types of tRNA that translate codon j. This amounts to the assumption that there is a single dominant step in the translational kinetics that is codon-dependent. The same assumption has been made in previous theories for codon usage (e.g. Shields, 1990; Bulmer, 1991; Solomovici *et al.* 1997). For simplicity we will assume that tRNA concentrations are proportional to tRNA gene numbers, which is approximately true experimentally (Kanaya *et al.* 1999), *i.e.* $C_i = c_0 N_i$, where N_i is the number of copies of the corresponding tRNA gene in the genome and c_0 is the concentration arising from one gene copy. The rate

constant for tRNA i with codon j will be written as $k_{ij} = k_0 b_{XY}$, where k_0 is an overall rate constant for translation, X is the base at the wobble position of the tRNA, Y is the base at the third position of the codon, and b_{XY} is a constant of order 1 that measures the relative rate of translation of the XY combination.

In general the rate of translation of codon i is the sum of the rates at which it is translated by all the tRNAs that interact with that codon:

$$r_i = c_0 k_0 \sum_j N_j b_{ji} . \tag{2.14}$$

The mean time to translate the codon is $1/r_i$. The selective advantage or disadvantage of codon i relative to the reference codon is assumed to be proportional to the difference in the times, Δt, between the two codons,

$$s_i = s_0 \Delta t = s_o (\frac{1}{r_{ref}} - \frac{1}{r_i}) . \tag{2.15}$$

where s_o is a genome-specific constant that determines the strength of selection in that genome. Note that s_i is positive if $r_i > r_{ref}$. We suppose that s_0 varies among genomes because the extent to which translational speed is important to an organism depends on its lifestyle, which can be dependent on growth rate difference. Organisms that multiply rapidly should be under significant selection to increase translational speed, and should have a high s_0.

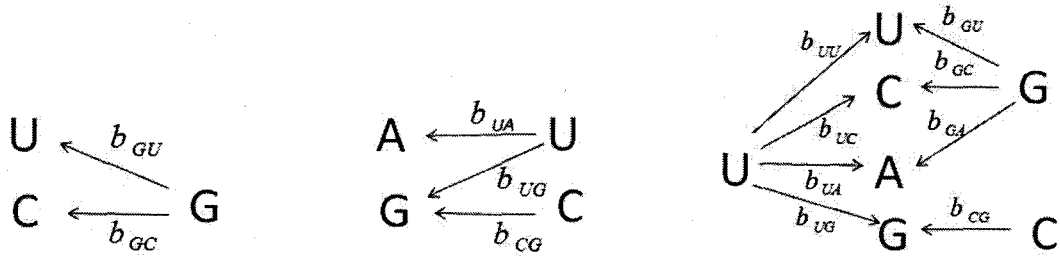
It will be useful to define relative rates ρ_i as

$$\rho_i = \frac{r_i}{c_0 k_0} = \sum_j N_j b_{ji} . \tag{2.16}$$

From this, the selection parameters S_i, which can be compared to the data, can be written as

$$S_i = 2N_e s_i = K(\frac{1}{\rho_{ref}} - \frac{1}{\rho_i}) \quad . \tag{2.17}$$

For convenience, we combined several parameters into a single parameter, $K = \frac{2N_e s_o}{c_o k_o}$.



a b c

Fig. 2.1 Translation kinetic model shows pairing possibilities for three types of codon family: (a) U+C families; (b) A+G families; (c) four-codon families. The arrow labeled b_{ji} indicates that a tRNA with wobble base j can pair with a codon with third position base i. The parameter b_{ji} denotes the relative rate of translating this codon by this type of tRNA. Only those combinations of anticodons and codons that are labelled by arrows are assumed to occur.

We will consider the three types of codon families described in section 1.2 and Figure 2.1. Only those combinations shown by an arrow in Figure 2.1 are assumed to be present in a cell. We can now write down the relative rates specifically for the three types of codon families.

U+C families –

$$\rho_U = N_G b_{GU}, \qquad \rho_C = N_G b_{GC}. \tag{2.18}$$

A+G families –

$$\rho_A = N_U b_{UA}, \qquad \rho_G = N_U b_{UG} + N_C b_{CG}. \tag{2.19}$$

Four-codon families –

$$\rho_U = N_U b_{UU} + N_G b_{GU}, \qquad \rho_C = N_U b_{UC} + N_G b_{GC},$$

$$\rho_A = N_U b_{UA} + N_G b_{GA}, \qquad \rho_G = N_U b_{UG} + N_C b_{CG}. \tag{2.20}$$

It should be remembered that the U and G bases at the wobble position are modified in many tRNAs and that this is likely to have a significant effect on the rates of interaction with the codons. We will consider these effects in detail later, but for simplicity of notation we use the unmodified form of the bases in the subscripts. Moreover, when analyzing the data, we wish to group cases with the same anticodons from a large number of different organisms, and the nature of the base modification is not known in every case. For this reason, we are obliged to consider only the unmodified form of the base at this stage.

The relative rates b_{XY} should be properties of the tRNA structures, the anticodon-codon interaction, and the way that recognition of the correct tRNA occurs in the ribosome. However, if the mechanism of translation is similar in different organisms, the relative rates should not depend on the organism (the tRNA structure and conformation are indeed different in different organisms. However, there is no good way to include that in our model and we assume that effect is same for all organisms). Then if a particular anticodon-codon combination leads to rapid translation in one organism, the same combination should function similarly in another organism. Thus we expect to see a

consistent relationship between preferred codons and tRNA genes that applies across many organisms.

In an U+C family, The relative translation rates are given by equation 2.18 and S is a function of N_G:

$$S(N_G) = \frac{K}{N_G}\left(\frac{1}{b_{GU}} - \frac{1}{b_{GC}}\right) \qquad (2.21)$$

If $b_{GC} > b_{GU}$, S will be positive in equation 2.21. In an A+G family, the selection coefficient is

$$S(N_U, N_C) = 2N_e s_0 (t_A - t_G) = K\left(\frac{1}{N_U b_{UA}} - \frac{1}{N_U b_{UG} + N_C b_{CG}}\right) \qquad (2.22)$$

This can be positive or negative, depending on the b parameters and the numbers of tRNA genes.

In equations 2.21 and 2.22 we have grouped the factors that affect the strength of selection into a single parameter K, and separated these from the factors that influence the direction of selection (*i.e.* the N and b parameters). One of our aims in Chapter 3 is to compare the cost of translation in different organisms. K is a useful parameter because it should be a property of an organism that depends on its lifestyle. The time of translation should be a significant fraction of the total cell division time. Hence in rapidly multiplying organisms, there should be significant selection to speed up translation and K should be large. In slowly multiplying organisms, the time for translation may not be a limiting factor. Hence, K may be small.

Chapter 3

Coevolution of Codon Usage and tRNA Gene Content

Preface

This theory shows that codon usage bias and tRNA content coevolve towards a stable state where each is adapted to the other. Multiple different stable states are found in many cases, which solves the puzzle that CUB can go in different directions in different amino acids in a same organism. The effect of effective population size on the variation of selection among different organisms is also discussed.

3.1 Introduction

As it is mentioned in section 1.1, synonymous codons in many species are not used with equal frequency. One of the major causes is potential selection for translational speed. In section 2.1 the population genetics theories have been introduced to calculate the way codon usage should depend on selection, mutation and drift (Li, 1987; Shields, 1990; Bulmer, 1991). In that section we also extended and improved these theories by taking more careful account of the way that the selective advantage of one codon over another should depend on the set of tRNA genes in the genome. This is a problem of co-evolution. For a given set of tRNA genes, codon usage will evolve to an equilibrium in which there is a bias towards codons that are rapidly translated by the current set of tRNAs. However, tRNA copy numbers can change due to gene deletions and duplications or anticodon mutations, as we show below. Therefore copy numbers can evolve to match the codon

usage. In general, we expect to find genomes in a stable co-evolved state in which neither codon usage nor tRNA gene content can change without decreasing the efficiency of translation. The theory that we give below predicts which combinations of codon usage and tRNA genes will be stable. The theory will be tested by comparison with observations in a large set of bacterial genomes.

As we saw in section 1.4 and Table 1.1, the smallest bacterial genomes have under 30 tRNA genes, all of which are distinct, whereas larger bacterial genomes can have over 120 tRNA genes, many of which are duplicate copies. Cell division times in bacteria differ from minutes to days. Rocha (2004) has shown that more rapidly multiplying bacteria tend to have larger numbers of tRNA gene copies. The likely explanation is that increasing the number of gene copies leads to increased tRNA concentration in the cell, which allows more rapid translation. Rapid translation is a significant advantage in rapidly multiplying organisms for which the time spent in translation is a significant limiting factor on the cell division time. In this chapter, we incorporate this qualitative observation into a quantitative theory. We show that in organisms for which the time cost of translation is significant (*i.e.* rapidly multiplying organisms) it is beneficial to duplicate tRNA genes. Selection on codon usage will also be stronger in these same organisms. Thus more strongly biased codon usage should occur in genomes with larger numbers of tRNAs.

The fact that tRNA copy numbers can vary by duplication and deletion is evident from the observation that different organisms have different total copy numbers. However, anticodon mutations are also important in tRNA evolution, because they can potentially

turn one type of tRNA into another. A good example of this is with tRNA Leu genes in mitochondrial genomes (Higgs *et al.* 2003) where several cases are known of interchange between genes with UAG and UAA anticodons. Lavrov and Lang (2005) have also found cases of anticodon mutations that convert a mitochondrial tRNA into a gene for a different amino acid. Anticodon mutations are also linked to several changes in the mitochondrial genetic code (Sengupta et al. 2007). The reassignment of the UGA stop codon to Trp occurs via a mutation in the anticodon of the Trp tRNA from CCA to UCA. The reassignment of AGY to Gly in urochordates occurs by duplication of a standard tRNA Gly with UCC anticodon, followed by mutation of one of the anticodons to UCU. Anticodon mutations have also been reported in bacterial tRNAs (Saks *et al.* 1998). When one compares bacterial genomes that are not too distantly related, it is often possible to find more than one distinct set of orthologous tRNA genes for a same amino acid. This suggests that anticodon substitutions are not too frequent. However, it is less clear whether this is true in more distantly related species because it becomes difficult to distinguish orthologs and paralogs in short tRNA sequences that contain a limited amount of phylogenetic information. The important point is that if a mutation occurs in an anticodon, the mutant sequence is likely to still be functional; therefore anticodon mutations are a potential means of evolution of the tRNA gene content of a genome. A recent study compared tRNAs in several complete genomes of *E. coli* and related species (Withers *et al.* 2006). They were able to distinguish a core set of tRNAs that has been present in these genomes for over a hundred million years from additional tRNAs that had been inserted by recent horizontal transfer in *E. coli* O157:H7 and *Shigella flexneri.*

In the following section, we use the population genetics theory in section 2.1 and show how this can be used to estimate the strength of selection acting on different codons. In the subsequent sections, we present a new theory that shows the way that translational kinetics and translational selection should depend on the number of tRNA gene copies. We show that co-evolution of tRNA genes and codon usage can lead to more than one possible stable state. This explains the observation that different amino acids in the same organism can have different preferred codons, which will be discussed in section 3.2. Finally, we carry out an analysis of a large number of bacterial genomes in order to test the predictions of the theory regarding the relationship between codon usage and tRNA copy number.

3.2 Examples showing the dependency of codon bias on tRNA gene copy number

Tables 3.1 and 3.2 show several examples of estimation selection strength from the sequences, which is the method introduced in section 2.1. In Table 3.1, we consider codons for Asn and Asp, two examples of U+C codon families, and in Table 3.2, we consider codons for Gln and Glu, two examples of A+G codon families. Four bacterial species and three eukaryotic were chosen as illustrative examples of the patterns of tRNA copy numbers that occur in the A+G families. The bacteria were chosen from among the 80 bacterial species previously analyzed by Sharp *et al.* (2005). The high expression gene set includes elongation factors Tu, Ts and G and 37 ribosomal proteins genes, as used by Sharp *et al.* (2005). For comparison, the tables also include three eukaryotes where translational selection is thought to be important. We calculated these data using the

Table 3.1 – Examples of codon usage, estimated selection strength, and tRNA gene copy numbers in representative species for two amino acids with U+C codon families.

		$\frac{n_C^{low}}{n_U^{low}+n_C^{low}}$	$\frac{n_C^{high}}{n_U^{high}+n_C^{high}}$	S	tRNA copies N_G
Mycoplasma penetrans	Asn	0.236	0.403	0.78	1
	Asp	0.140	0.201	0.43	1
Agrobacterium tumefaciens	Asn	0.561	0.841	1.42	1
	Asp	0.491	0.764	1.21	2
Lactobacillus plantarum	Asn	0.397	0.774	1.65	5
	Asp	0.340	0.458	0.49	3
Escherichia coli	Asn	0.550	0.875	1.74	4
	Asp	0.372	0.657	1.17	3
Schizosaccharomyces pombe	Asn	0.343	0.718	1.58	6
	Asp	0.292	0.438	0.64	8
Saccharomyces cerevisiae	Asn	0.410	0.860	2.18	10
	Asp	0.350	0.578	0.93	16
Caenorhabditis elegans	Asn	0.378	0.575	0.80	20
	Asp	0.324	0.559	0.97	27

Table 3.2 – Examples of codon usage, estimated selection strength, and tRNA gene copy numbers in representative species for two amino acids with A+G codon families.

		$\frac{n_G^{low}}{n_A^{low}+n_G^{low}}$	$\frac{n_G^{high}}{n_A^{high}+n_G^{high}}$	S	tRNA copies N_U:N_C
Mycoplasma penetrans	Gln	0.063	0.017	-1.33	1:0
	Glu	0.094	0.033	-1.11	1:0
Agrobacterium tumefaciens	Gln	0.831	0.982	2.40	1:1
	Glu	0.427	0.273	-0.69	2:0
Lactobacillus plantarum	Gln	0.360	0.046	-2.45	2:1
	Glu	0.245	0.030	-2.35	2:1
Escherichia coli	Gln	0.653	0.813	0.84	2:2
	Glu	0.311	0.244	-0.34	4:0
Schizosaccharomyces pombe	Gln	0.285	0.156	-0.77	4:2
	Glu	0.322	0.565	1.01	4:6
Saccharomyces cerevisiae	Gln	0.307	0.009	-3.89	9:1
	Glu	0.300	0.028	-2.70	14:2
Caenorhabditis elegans	Gln	0.343	0.307	-0.16	20:7
	Glu	0.375	0.436	0.25	17:24

sequences of the ribosomal proteins obtained from the ribosomal protein gene database (Nakao *et al.* 2004) and the sequences of three elongation factors from NCBI database.

In Table 3.1, N_G denotes the number of tRNA gene copies for the amino acid, all of which have G at the wobble position. In Table 3.2, N_U and N_C denote the number of tRNA gene copies for the amino acid that have U or C at the wobble position. In Table 3.1, we see that S is positive for Asn and Asp in each species. This agrees with our expectations because the wobble G pairs better with C. In Table 3.2, we see examples of both positive and negative S for Gln and Glu, and the direction of selection depends on the tRNA copy numbers. Cases where $N_C = 0$ (*ie.* 1:0, 2:0, 4:0) all have negative S. This shows that when the only tRNA has wobble-U, it interacts more efficiently with the A codon, and the A codon is preferred. In cases where $N_U > N_C > 0$ (such as 2:1 and 4:2) S is still negative and A is still preferred. In cases where $N_U = N_C$, or $N_U < N_C$, S is positive and G is preferred. This shows that if there are a sufficient number of wobble-C tRNAs relative to wobble-U tRNAs, the direction of selection on codon usage is reversed. This agrees with our expectation that wobble-C tRNAs interact with G codons but not A codons.

These examples are intended as motivation for the theory that follows. The bacterial species chosen are illustrative of the trends in the larger data set of 80 bacterial species that we considered and many other examples could have been chosen. We will give a statistical analysis of the full data set after presenting the theory. The three eukaryotes are included because these species are among the eukaryotes whose codon usage has been studied in most detail. The theory below is intended principally as a

theory of codon usage in bacteria, but it also applies well to these three eukaryotes. For other eukaryotes such as humans and *Drosophila*, codon usage does not seem to be dominated by translational selection in the same way. Thus we will not discuss eukaryotes further in this thesis.

An interesting feature of Table 3.2 is that there are several species where selection goes in opposite directions for Gln and Glu, *i.e.* where G is preferred for one amino acid and A is preferred for the other. The coevolution of codon usage and tRNA gene content has led to different stable states in the same organism. We now wish to show how this situation can arise using a simple model for translation kinetics.

3. 3 Coevolution with fixed total tRNA copy number

In this section we will consider an A+G codon family with tRNA copy numbers N_U and N_C. We suppose that the total number of tRNAs is fixed, but that the number of copies of each type can vary. If the total number of copies is four, as is the case for Gln and Glu in *E. coli* in Table 3.2, then the possible combinations of N_U:N_C are 4:0, 3:1, 2:2 and 1:3. The combination 0:4 is forbidden because there must be at least one U tRNA to translate the A codons. The situation of fixed total copy number is discussed first because it is the simplest. We generalize to variable total number in the following section. However, the fixed total copy number case is similar to the situation considered by Bulmer (1987) and Shields (1990), who suppose that there are two tRNAs with fixed total concentration. We will discuss the differences that arise between our theory and the previous one at the end of this section.

Let ϕ^* denote the frequency of the G codon in an A+G family that would arise under SMD with specified tRNA copy numbers, *i.e.* $\phi^*(N_U, N_C) = \phi(S(N_U, N_C))$, which can be calculated from Equations 2.13 and 2.22. Figure 3.1(a) shows ϕ^* as a function of K for each possible combination of N_U:N_C. As the b parameters are relative rates, we can set one of them to 1 by definition. We will use the U tRNA + A codon combination as a reference, *i.e.* $b_{UA} = 1$. In the examples used in this thesis we will also set $b_{CG} = 1$, for simplicity since both are Watson-Crick pairs. Based on strengths of RNA base pair interaction, we would expect that $b_{UG} < 1$, because UG pairs are weaker than Watson-Crick pairs. We will set $b_{UG} = 0.4$ in these examples. We will show below that these parameter values appear to be close to optimal for interpreting the codon usage on bacterial genomes. For these parameter choices, $S(4,0)$ and $S(3,1)$ are negative, but $S(2,2)$ and $S(1,3)$ are positive; therefore $\phi^*(4,0)$ and $\phi^*(3,1)$ decrease with K, but $\phi^*(2,2)$ and $\phi^*(1,3)$ increase with K, as shown in Figure 3.1(a).

In genes where the frequencies of G and A codons are ϕ and 1-ϕ, the mean time per codon is

$$\bar{t}(\phi, N_U, N_C) = \phi t_G + (1-\phi)t_A = \frac{1}{k_0 c_0}\left(\frac{\phi}{N_U b_{UG} + N_C b_{CG}} + \frac{1-\phi}{N_U}\right). \qquad (3.1)$$

Figure 3.1(b) shows $\bar{t}$ as a function of ϕ for each combination of tRNAs. We have set $k_0 c_0 = 1$ in the figure because this is simply a multiplying factor. Selection is acting to minimize $\bar{t}$. The gradients of these lines are proportional to the selection coefficients. The 4:0 and 3:1 lines slope down towards $\phi = 0$, corresponding to negative S, and the 2:2 and 1:3 lines slope down towards $\phi = 1$, corresponding to positive S. The codon frequency

that would optimize translation time is always either 0 or 1, depending on the direction of the slope. However, the frequency that occurs is ϕ^*, which is between 0 and 1 because it is also influenced by mutation and drift, not just by selection.

In a real genome, the anticodons of the tRNAs can adapt to the codon usage at the same time as the codon usage adapts to the tRNAs. A straightforward mutation at the wobble position can convert a U tRNA to a C tRNA, as discussed in the introduction. Thus an organism can jump between the lines on Figure 3.1(b). If the codon frequencies are at SMD balance with specified N_U and N_C, the mean time per codon is $\bar{t}(\phi^*(N_U,N_C),N_U,N_C)$. If this combination of tRNAs is stable to anticodon mutation, then this time must be less than the time that would occur at the same codon frequencies with any other copy numbers M_U and M_C such that $M_U+M_C = N_U+N_C$:

$$\bar{t}(\phi^*(N_U,N_C),N_U,N_C) < \bar{t}(\phi^*(N_U,N_C),M_U,M_C). \tag{3.2}$$

In other words, if a tRNA gene combination is stable, then the $\bar{t}$ line for this combination must be the lowest line of those on Figure 3.1(b). For each of the N_U:N_C combinations, there is a range of ϕ where the corresponding time is the lowest. The boundaries between these regions are indicated by dotted vertical lines in Figure 3. 1(b). The same boundaries are indicated by dotted horizontal lines in Figure 3.1(a). In order for the N_U:N_C combination to be stable to anticodon mutations, $\phi^*(N_U,N_C)$ must lie within the corresponding boundaries. Stable regions are indicated by thick black lines in Figure 3.1(a). In these regions, the tRNAs and codons are co-adapted. In the regions of the curves drawn within lines, the codons are adapted to the tRNAs, but the tRNAs are not adapted to the codons. Hence, the configuration is not stable to anticodon mutations.

Fig 3.1

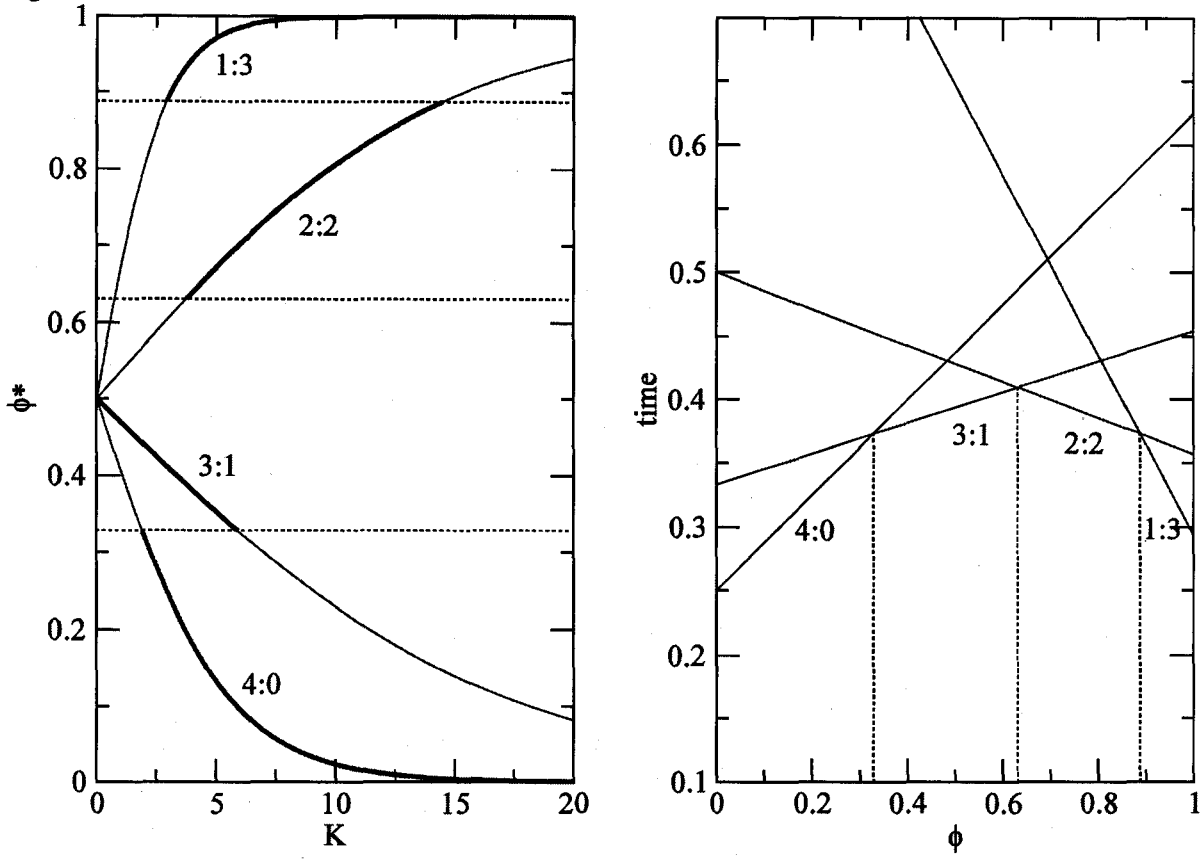


Fig. 3.1. (a) Expected frequency ϕ of the G codon in A+G families as a function of translational cost parameter K. Four different combinations of tRNA gene copy numbers N_U:N_C are shown. Thick black lines indicate regions of stability. (b) Mean translation time per codon as a function of ϕ for each tRNA gene combination.

For the example illustrated in Figure 3.1 the GC content arising from mutation is $\theta = 0.5$. Thus all curves tend to 0.5 as K tends to zero. The line lies in the interval where the 3:1 combination is stable. If K increases, the other three combinations become stable within certain intervals of K. For moderate or large K there is always more than one stable solution. This is the central point of this chapter. In species where there is significant translational selection, alternative stable states of codon usage exist where codons and tRNAs are co-adapted. It is therefore possible for the codon usage in codon families for different amino acids to be biased in different directions, even if they are subject to the same mutation and selection processes, as we saw in the examples in Table 3.2.

In this theory, we are treating mutations in anticodons in a different way from those in codons. For synonymous mutations in codons, we assume that selection is balanced by mutation and drift, so that at any synonymous site it is possible for either an optimal or non-optimal codon to occur with some probability. In contrast, if a favourable mutation occurs at the wobble position of the anticodon, we assume that it is always selected. This is because a mutation at a synonymous site affects the translation time of only one codon, whereas a mutation in the anticodon affects the translation time of all the codons for that amino acid. Selection on the wobble position is therefore orders of magnitude stronger than on a synonymous site. If selection is large enough to cause a significant bias at synonymous sites, then selection on the wobble position should be very large indeed, and it is reasonable to assume that the optimal state of the anticodons will always be fixed in the population. This point has also been made by Jia and Higgs (2008) with regard to the evolution of tRNAs and codon usage in mitochondrial genomes.

As we have assumed that the concentration of tRNAs is proportional to the number of gene copies and that the total number of gene copies is fixed, this case is similar to previous theories that considered concentrations of two tRNAs with a fixed total concentration, $C_1 + C_2 = 1$. However, there are important differences. The selection parameter used by Bulmer (1987) and Shields (1990) is $s = k(1/C_2 - 1/C_1)$, which may be compared with our equation 2.22. This has the problem that it becomes infinite if either C_1 or C_2 goes to zero. It does not capture the way anticodon-codon interactions work because it assumes that two separate tRNAs translate the two different codons, whereas in reality, wobble G tRNAs can translate both codons in U+C families, and wobble-U tRNAs can translate both codons in A+G families. In A+G families, it is possible for N_C to be zero when N_U is non-zero, because the U tRNA translates both codons. The selection strength in equation 2.22 can never be infinite in this case. On the other hand, N_U cannot be zero because then it would be impossible to translate A codons. Shields (1990) found that, if tRNA concentrations were treated as continuous variables, there would be a synergistic increase of tRNA concentrations and codon frequences leading to ever increasing biases. He thus concluded that organisms would exhibit either no codon bias or complete codon bias. We disagree with this. The bias is always finite according to our theory because the selection parameter is always finite.

Interestingly, the problem of ever increasing biases in the Shields theory applies only to the case where all genes are assumed to have equal expression levels. He also considered a model with separate classes of high and low expression genes, assuming that a small proportion of highly expressed genes accounts for half the total gene expression,

and assuming a lower selection strength in the low-expression genes. In our theory, we have made the more extreme assumption that the translational effort of the cell is dominated by the high expression genes. Therefore only the time $\bar{t}$ for the high-expression genes is relevant for the translational cost, and we only need to consider selection on one class of genes. It would be possible to modify our theory to include a reduced level of selection on low-expression genes and a parameter to control the fraction of the total translational effort of the cell that goes into high- and low-expression genes. However, we do not want to make the theory more complicated with additional parameters at this point. We emphasize that it is not necessary to consider genes with two different selection levels in order to avoid the problem of ever-increasing biases. The problem is avoided by a more realistic choice of the selection parameter in our model.

3. 4 Coevolution with variable total tRNA copy number

In reality, tRNA copy numbers can change by gene duplication and deletion as well as anticodon mutation. Therefore, the total tRNA copy number for an amino acid is not fixed. In organisms where translational selection is strong, duplicate copies of a tRNA gene will be made so that the tRNA concentration will increase and translation will be faster. However, there is some cost to the organism for duplicating this gene. Bacteria do not often retain redundant duplicate genes because they are under selection for rapid replication and therefore their genome size tends to be minimized. Duplicating the tRNA therefore has a cost in terms of increased DNA replication time. It also has a cost in terms of transcription. When the gene is present, the organism will expend time and energy in

making tRNA molecules by transcribing this gene. This would be disadvantageous to the organism if the extra tRNAs were not beneficial for translation. We expect the total cost of translation of codons of a given amino acid to be

$$T = 2N_e s_o f_a \bar{t} + gN_a \tag{3.3}$$

where f_a is the frequency of codons for the amino acid in high expression genes, N_a is the number of tRNA genes for the amino acid and g is the duplicate time of a tRNA gene.

The U+C families become of greater interest when the number of tRNAs can vary. For a U+C family, the translational cost T is a function of the frequency ϕ of the C codon, and the number of wobble-G tRNAs. From equations 2.17, 2.18 and 3.3, we have

$$T(\phi, N_G) = \frac{f_a K}{N_G}\left(\frac{\phi}{b_{GC}} + \frac{1-\phi}{b_{GU}}\right) + gN_G \tag{3.4}$$

At the SMD equilibrium for a fixed N_G, the codon frequency is $\phi^*(N_G) = \phi(S(N_G))$ from Equations 2.13, 2.17 and 2.18. If this solution is stable against duplication or deletion of tRNAs, we must have

$$T(\phi^*(N_G), N_G) < T(\phi^*(N_G), M_G) \quad \text{for any } M_G \neq N_G. \tag{3.5}$$

Figure 3.2 shows the ϕ^* curves as a function of K for each value of N_G. Regions that are stable according to Equation 3.5 are indicated by thick black lines. In this example, b_{UA} = b_{GC} = 1, b_{GU} = 0.4, and θ = 0.5. If K is small enough, T is minimized by setting N_G = 1, irrespective of the other parameters. Above a certain value of K, the N_G = 1 solution becomes unstable. As K increases, each successive value of N_G becomes stable in a range that partially overlaps the previous value of N_G, so once again, there can be more than one stable state for a given K. The ranges of ϕ covered by the stable regions of the curves are

Fig 3.2

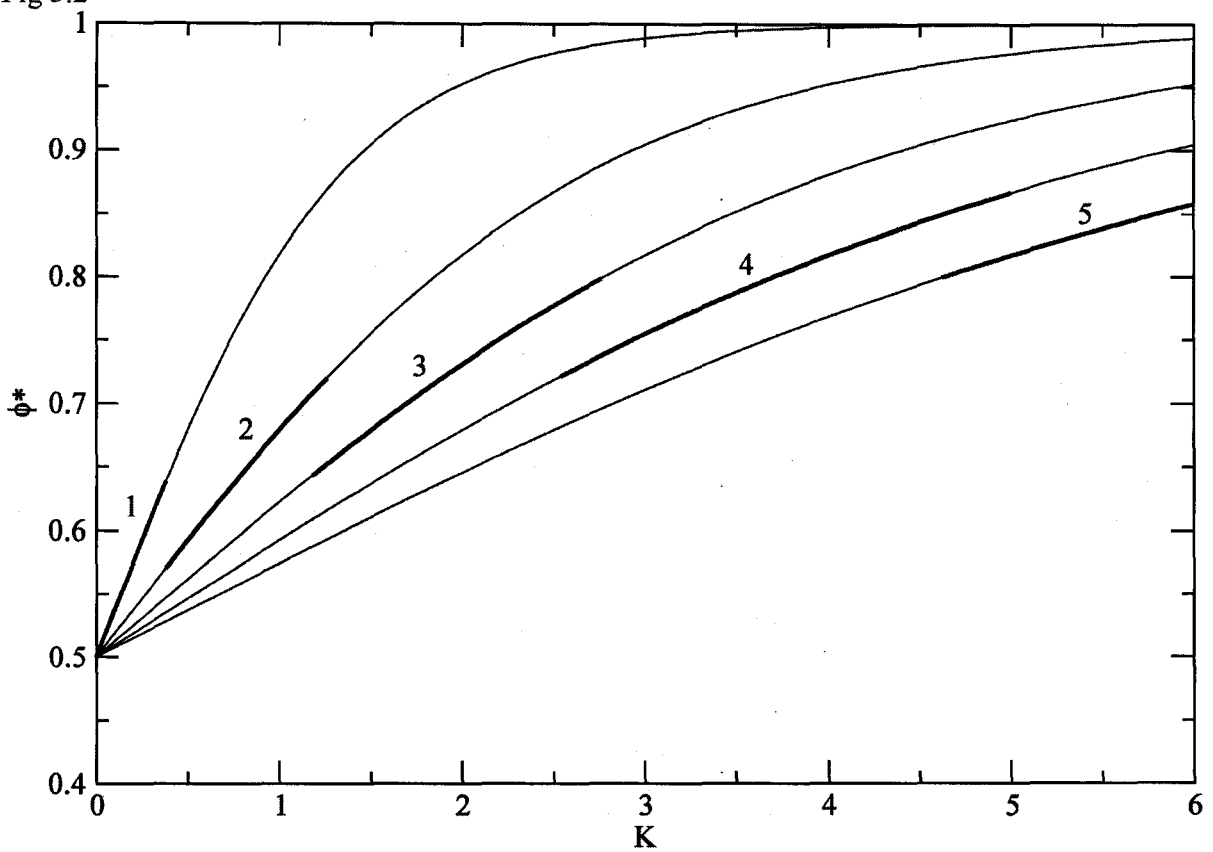


Fig. 3.2. Frequency of the C codon in U+C families as a function of K for varying numbers N_G of tRNA copies. Thick black lines show regions of stability by equation 3.5.

quite broad. Thus there might be a considerable amount of variation in the degree of codon bias observed in organisms with the same number of tRNA genes. Paradoxically, because the strength of selection varies inversely with N_G, the degree of codon bias actually decreases after a tRNA gene duplication. However, the stable range of ϕ shifts towards more biased codon usage as N_G increases; therefore, on average, we expect to see stronger codon bias in organisms with more tRNA genes.

The regions of stability depend on the ratio g/f_a, which is set to 0.3 in this example. If this ratio is lower, the regions of stability slide down the curves to lower ranges of K, although the positions of the curves themselves do not depend on g/f_a. Thus it is more favourable to add an extra tRNA if the cost per tRNA gene is lower or if the frequency of the amino acid is higher.

We now return to A+G codon families and consider both anticodon mutations and tRNA duplications and deletions at the same time. In the same way as above, the translational cost is

$$T(\phi, N_U, N_C) = f_a K\left(\frac{\phi}{N_U b_{UG} + N_C b_{CG}} + \frac{1-\phi}{N_U}\right) + g(N_U + N_C). \tag{3.6}$$

The stability criterion is

$$T(\phi^*(N_U, N_C), N_U, N_C) < T(\phi^*(N_U, N_C), M_U, M_C), \tag{3.7}$$

and this must apply for all combinations $M_U{:}M_C \neq N_U{:}N_C$.

Figure 3. 3 shows an example with $b_{UA} = b_{CG} = 1$, $b_{UG} = 0.4$, $\theta = 0.5$, and $g/f_a = 0.3$. For clarity, only the stable regions of the curves are shown. As with Figure 3. 2, the solution with only one tRNA is stable at very low K because the time dependent term in

the cost function becomes smaller than the term involving the cost per gene, so the total cost is minimized by minimizing the number of genes. Solutions with larger numbers of tRNAs become more stable as K increases. This is different from the situation in Figure 3. 1, where the 3:1 solution was stable at low K. The parameters in Figures 3.1 and 3.3 are the same, with the exception of the addition of g/f_a in Figure 3. 3. The 3:1 solution is stable against anticodon mutations at low K but not against tRNA deletion.

There are some combinations (*e.g.* 1:2, 2:2, 3:2 and 2:3) that do not appear on Figure 3.3 because there is no stable region of the corresponding curve for any K. The combinations that have a stable range depend on the other parameters, and in particular, they are influenced by θ, the GC content specified by the mutation rates. In Figure 3.4, θ = 0.7, rather that 0.5, as it was in Figure 3. 3. In this case 1:2, 2:2, 3:2 and 2:3 all have a stable region, but there is no stable region for 2:0, 3:0 and 4:0.

As θ has an important influence on the stability of the different solutions, it is interesting to consider the way codon frequencies and tRNA combinations are likely to vary with θ when K is fixed. Figure 3. 5 shows two examples with K = 0.5 and 3.0. The other parameters are as before. For K = 0.5, only low tRNA number solutions are stable (1:0, 2:0 and 1:1). Each of these is stable within an interval of θ. As selection is weak in this example, ϕ does not differ much from θ on any of these curves (all the curves lie close to the diagonal). For K = 3.0, solutions with larger tRNA numbers are stable (3:0, 4:0, 3:1, 2:2 and 1:3). There is significant codon bias in this example: ϕ can be considerably higher or lower than θ, especially for the more uneven tRNA combinations.

Fig 3. 3

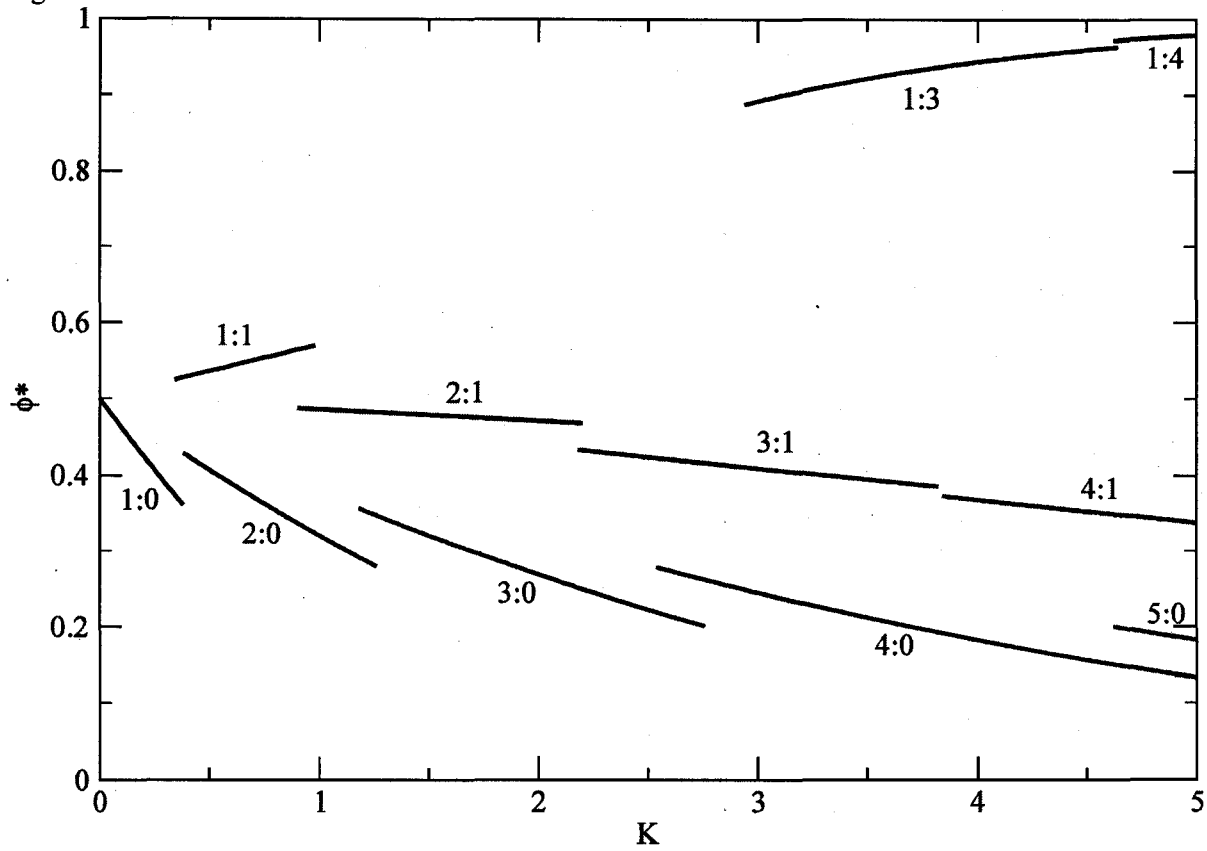


Fig. 3.3. Stable solutions for codon frequency as a function of K in A+G families where the GC content of the mutational process is $\theta = 0.5$. Labels indicate tRNA gene numbers N_U:N_C.

Fig 3. 4

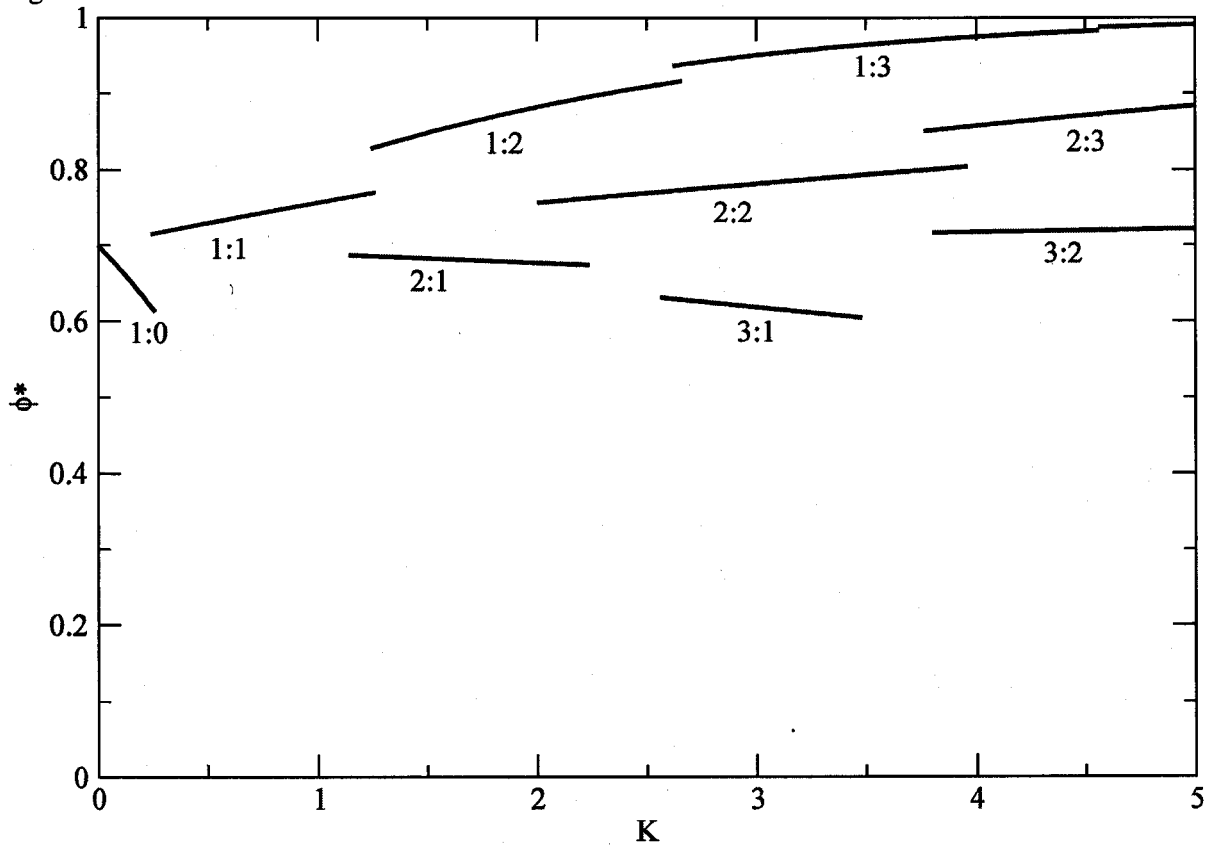


Fig. 3.4. As Figure 3.3 except that $\theta = 0.7$.

Figure 3.5 gives an idea of what might be expected to happen if there is a gradual change in the GC content of the genome of an organism with time due to changing of the relative mutation rates between the different bases. A GC rich organism with $\theta = 0.9$ might initially lie on the 1:3 curve. If θ is gradually reduced, the GC content in the majority of genes in the genome will follow this, but the GC content in the high expression genes will follow the 1:3 curve and will therefore remain very high until this curve becomes unstable at θ close to 0.5. At this point there will be a shift in the tRNA content and a sudden change in the codon usage in the highly expressed genes. A similar behaviour would also occur according to the theory of Shields (1990).

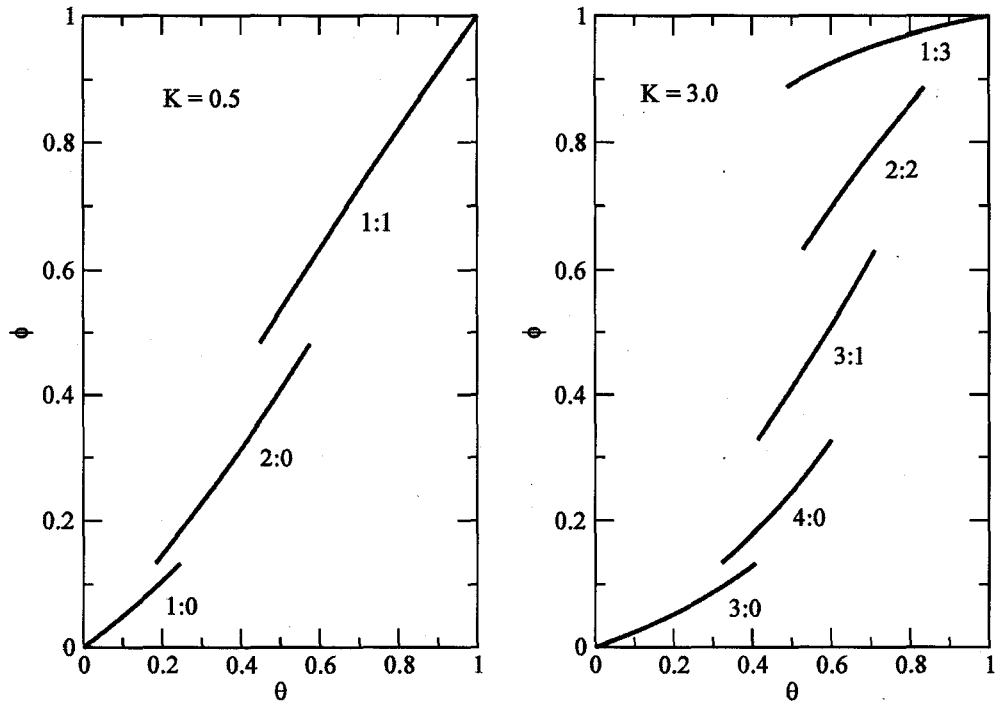


Fig. 3.5. Variation of codon usage ϕ with the GC content of the mutation process θ shown for two different values of translational cost parameter K. Stable solutions with different tRNA gene combinations are shown.

3. 5 Comparison of theory with bacterial codon usage data

We now return to the analysis of the 80 bacterial genomes. Sharp *et al.* (2005) calculated an average of S in U+C amino acids. As discussed above, the cost of translation is determined by K. Although the strength of selection S is proportional to K, it is better to average K over amino acids than to average S, because S is affected by the number of tRNA gene copies, which is variable. For a U+C amino acid a, the estimated selection strength is S_a (determined from equation 2.15), and the estimated cost parameter is $K_a = N_G S_a /(1/b_{GU} - 1/b_{GC})$, where N_G is known from the genome. The factor involving the b parameters is simply a multiplying factor that we assume is the same for each amino acid. For consistency with the examples above, we have used $b_{UA} = b_{GC} = 1$ and $b_{GU} = 0.4$, but the relative values of K for each organism do not depend on this choice of b parameters. Thus, as a measure of the cost of translation in an organism, we will use

$$\overline{K} = \left(\sum_a f_a K_a\right) \Big/ \left(\sum_a f_a\right). \tag{3.8}$$

where f_a is the frequency of amino acid a in the high expression protein sequences. Sharpe *et al.* used the four amino acids Phe, Ile, Tyr and Asn in their average. We will also include His and Asp, making six U+C amino acids in total, because these two behave the same way as the other four. We do not include Cys and Ser(AGY) because there are usually fewer codons for these amino acids and the statistics are less reliable.

Rocha (2004) estimated the minimum doubling times for different bacteria. He showed that the number of tRNA genes in the genome decreases with doubling time and increases with codon bias (estimated from comparing effective numbers of codons in high

and low expression genes). Our theory explains why this occurs. Our estimate of $\overline{K}$ is a measure of the time-dependent cost of translation, and hence of the benefit to be gained by optimizing translation. Figure 3.6(a) shows the relationship between doubling time and $\overline{K}$ for 77 of the 80 species considered here. Three species where no doubling time estimate was available were excluded (*Aquifex aeolicus*, *Mycoplasma penetrans*, *Wigglesworthia glossinidia*). There is a strong negative correlation between doubling time and $\overline{K}$ because translational selection is strongest in fast growing organisms that need to synthesize proteins very rapidly. Figure 3.6(b) shows that the total number of tRNA genes increases as a function of $\overline{K}$. Our theory explains why tRNA duplication is favourable in species with high K, and therefore explains why tRNA numbers increase with K and decrease with doubling time.

In contrast to this, the diversity of tRNAs (the number of distinct anticodons) appears to decrease slightly with $\overline{K}$, as shown in Figure 3.6(b). A similar point was noted by Rocha (2004), who showed that tRNA diversity is lower in the fast growing species than the slow growing ones. A possible explanation of this is that when translational selection is strong, codon bias is strong, and so it is possible for the organism to optimize its tRNAs by specializing to the preferred codons, and deleting the tRNAs that would best match the rare codons. To consider this argument more fully, it will be necessary to extend our theory to deal with tRNAs in four-codon families, which we will do in next chapter. One point that is apparent from the theory of A+G codon families above is that tRNA diversity will be influenced substantially by GC content. When θ is small and K is large, tRNA combinations such as 3:0 and 4:0 are expected to be stable (Figure 3.5),

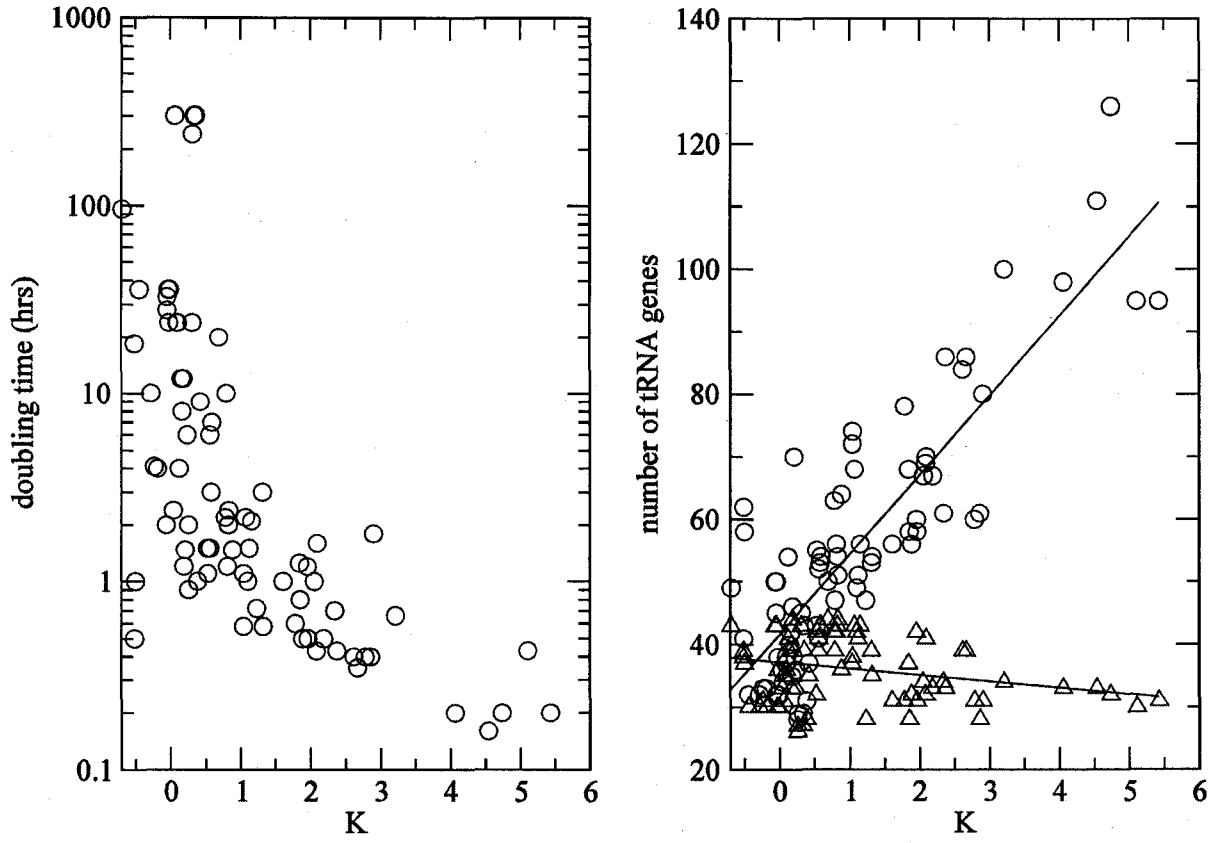


Fig. 3.6. (a) Relationship between the minimum doubling time of bacteria and the estimated value of the translational cost parameter $\overline{K}$. (b) Dependence of the total number of tRNA genes in bacterial genomes (circles) and the total number of different anticodons among the tRNA genes (triangles) on the estimated $\overline{K}$.

whereas when θ is large and K is large, combinations like 1:3 will be stable, which requires two distinct tRNAs instead of one. Thus, we expect that tRNA diversity will be larger in high GC organisms. In fact, it has already been shown that this is the case in bacterial genomes – see Figure 4D of Kanaya *et al.* (1999). Although the downward trend of tRNA diversity with $\overline{K}$ appears to be significant in Figure 3.6(b), this result should be treated with caution until a more complete analysis of four-codon families has been made.

We will now consider the way the observed tRNA gene combinations depend on $\overline{K}$. There are 14 of the 80 bacterial species for which $\overline{K}$ is negative: *Borrelia burgdorferi, Buchnera aphidicola (Ap, Bp and Sg), Chlamydophila pneumoniae, Chlorobium tepidum, Neisseria meningitidis, Nitrosomonas europaea, Pseudomonas aeruginosa, Rickettsia conorii, R. prowazeckii, Treponema pallidum, Tropheryma whipplei* and *Xylella fastidiosa*). With the exception of *C. tepidum*, these species also have negative $\overline{S}$. Sharp *et al.* (2005) give two possible explanations of negative $\overline{S}$ values: either base composition may be skewed between strands and high expression genes may be predominantly on the leading strand, or there may be islands of unusual base composition arising from horizontal transfer. These species are almost all slow growing organisms with low tRNA numbers. It is likely that translational selection is very weak in these species and that the small effects mentioned above are more important than translational selection. As tRNA-dependent translational selection is not the dominant effect in these organisms, they are a poor test of this theory; therefore we exclude these species from the following analysis. The remaining 66 species cover a wide range of $\overline{K}$, doubling time and total tRNA number and provide a good test set.

Table 3.3 summarizes information from six U+C amino acids (Phe, Ile, Tyr, His, Asn, Asp) in the 66 species. Table 3.4 summarizes information from three A+G amino acids (Gln, Lys, Glu) in the same species. Each row is a category corresponding to a given number of tRNA genes. N_{obs} is the number of observations in each category. The 'mean K' is the mean value of $\overline{K}$ for all the observations in each category. In Table 3.3, the mean K increases with G-tRNA copy number N_G. In Table 3.4, the mean K increases with U-tRNA copy number N_U in the categories from 1:0 to 7:0, and also from 1:1 to 3:1, and from 1:2 to 7:2. These results confirm the prediction of the above theory that larger numbers of tRNAs should be found in organisms with higher K. Note that only the U+C amino acids were used in order to estimate $\overline{K}$ for each species. However, these results show that $\overline{K}$ is also a predictor of what happens in the A+G amino acids. If selection is strong in an organism it causes higher numbers of tRNAs to arise for both U+C and A+G amino acids.

We also wish to test the sign of the selective effect. The tables show the times of translation of the two codons (with k_0c_0 factor set to 1) and the difference in the times. The sign of the selective effect should be the same as the sign of the time difference. The mean S column of the tables is the mean of S_a for each of the observations in the category. In Table 3.3, the mean S is positive in every case, as expected. In Table 3.4, the mean S can be either positive or negative, but it has the same sign as the time difference in every case except for the 7:0 category, in which there are only two observations. We also calculated N_{sign}, the number of observations for which S_a has the same sign as the time difference. In Table 3.3, almost all the observations in each category have the correct sign.

Table 3.3 Comparison of theory for U+C codon families with observations in bacteria.

NG	tU	tC	tU-tC	Nobs	mean K	mean S	Nsign
1	2.50	1.00	1.50	207	0.61	0.75	193
2	1.25	0.50	0.75	94	1.82	1.19	94
3	0.83	0.33	0.50	57	2.08	1.03	56
4	0.63	0.25	0.38	23	3.38	1.67	23
5	0.50	0.20	0.30	10	3.00	1.37	9
6	0.42	0.17	0.25	3	4.80	1.07	3
7	0.36	0.14	0.21	1	4.74	3.10	1

Table 3.4 Comparison of theory for A+G codon families with observations in bacteria.

NU	NC	tA	tG	tA-tG	Nobs	mean K	mean S	Nsign
1	0	1.00	2.50	-1.50	46	0.32	-0.26	31
2	0	0.50	1.25	-0.75	21	1.80	-1.66	20
3	0	0.33	0.83	-0.50	16	2.21	-0.55	12
4	0	0.25	0.63	-0.38	15	2.85	-0.64	14
5	0	0.20	0.50	-0.30	8	3.35	-0.42	6
6	0	0.17	0.42	-0.25	5	3.55	-0.21	4
7	0	0.14	0.36	-0.21	2	4.16	0.07	1
1	1	1.00	0.71	0.29	47	0.64	0.73	37
2	1	0.50	0.56	-0.06	14	1.34	-0.40	8
3	1	0.33	0.46	-0.12	7	1.97	-0.64	5
1	2	1.00	0.42	0.58	5	1.32	2.68	5
2	2	0.50	0.36	0.14	3	1.83	0.48	3
4	2	0.25	0.28	-0.03	1	1.04	-0.23	1
7	2	0.14	0.21	-0.07	1	5.43	-0.19	1

In Table 3.4, a majority of observations have the correct sign, but by no means all. One reason for this is probably statistical, because S_a is estimated from relatively small numbers of codons in the high frequency genes. The statistical error will be more likely in categories where the true selective effect is smaller. The true effect is small when K is small, such as the 1:0 category, or for categories that are close to the balance point where the two codons have equal times, such as the 2:1 category.

It should be remembered that the predicted sign of the effect depends on the b parameters. Throughout this chapter, we have assumed Watson-Crick pairs are stronger than non-Watson-Crick pairs. Then we have $b_{UA} = b_{CG} = b_{GC} = 1$, and $b_{UG} = b_{GU} = 0.4$. If the parameters are shifted too much from these values, the agreement with the observations becomes worse. For example, there is a clear majority of observations with positive selection in the 1:1 category. Thus from equation 2.20, we have the inequality $b_{UG} + b_{CG} > b_{UA}$. There is also a clear majority of observations with negative selection in the 3:1 category; hence $3b_{UG} + b_{CG} < 3b_{UA}$. The observations in the 2:1 category are more evenly split, which suggests that there is little difference in the times for the two codons when there is a 2:1 ratio of tRNAs.

One of the reasons the sign is not correctly predicted in 100% of the cases may be that the rates depend on the other two positions in the anticodon, not just the wobble position. In that case the relative rates of translating the two codons may not be the same for each amino acid, so no single set of b parameters would make correct predictions for Gln, Lys and Glu at the same time. It is also possible that there is variation in the level of transcription arising from different tRNA gene copies, so that the tRNA concentrations

are not exactly proportional to the gene copy numbers. This could also lead to a shift in the expected sign of the selective effect. Furthermore, it should be remembered that the wobble positions of some tRNAs are changed to modified bases. We have referred to the unmodified bases only, because the modifications are not known directly from the tRNA gene sequences. However, if different modifications occur in different organisms or in different tRNAs in the same organism, this could also change the b parameters.

Another prediction of our theory is that the tRNAs that are most likely to be duplicated in larger genomes are those whose corresponding codons are most frequent. Figure 3.7 compares the number of tRNA genes in genomes with high and low total tRNA numbers. Of the 80 species considered, the 10 genomes with 80 or more tRNAs were included in the high set. The 20 genomes with 40 or fewer tRNAs were included in the low set. The mean number of tRNAs per codon family was calculated for each set and this is shown as a function of the mean frequency of the corresponding codons in high expression genes in the genomes. U+C, A+G and 4-codon families are labelled by different symbols in the figures. The amino acids with 6 codons (Leu, Ser, Arg) are treated as two separate families of two and four codons because the tRNAs are easily distinguishable between these codons. The UGG Trp codon is a single codon family. The AUG Met codon is also a single codon, however the anticodon for tRNA-Met is CAU, and this is not easily distinguishable from the tRNA-Ile that pairs with the AUA codon. The latter has anticodon K_2CAU, where K_2C is lysidine. This post-transcriptional modification is not easily apparent from the gene sequence. Therefore we treat AUA and

AUG as a two-codon family and include all tRNAs whose unmodified anticodon is CAU in this family. The other Ile codons (AUU and AUC) are treated as a U+C family.

It can be seen that there is a clear correlation between tRNA number and codon frequency in the high set but not in the low. In genomes with low total number of tRNAs, there is almost always 1 tRNA for every U+C family. For A+G families the mean number is between 1 and 2 because there is occasionally a wobble-C tRNA in addition to a wobble-U. For 4-codon families the mean number is around 2 (usually a wobble-U and a wobble-G). For the AUA/G family the mean number is close to 3 because there is almost always an elongator and an initiator tRNA for Met and a tRNA-Ile with a lysidine modification. The AUA/G family is also an outlier in the high total number set, presumably because all these types of tRNA are duplicated. Genomes in the high set correspond to those with strong translational selection, where translational cost is minimized by duplication of tRNAs in proportion to the frequency of their codons. Genomes in the low set correspond to those where translational selection is weak and it does not pay to duplicate genes. In these genomes the main constrain on the tRNA set is that it must remain diverse enough to translate the full set of codons, but the frequency of these codons is not important.

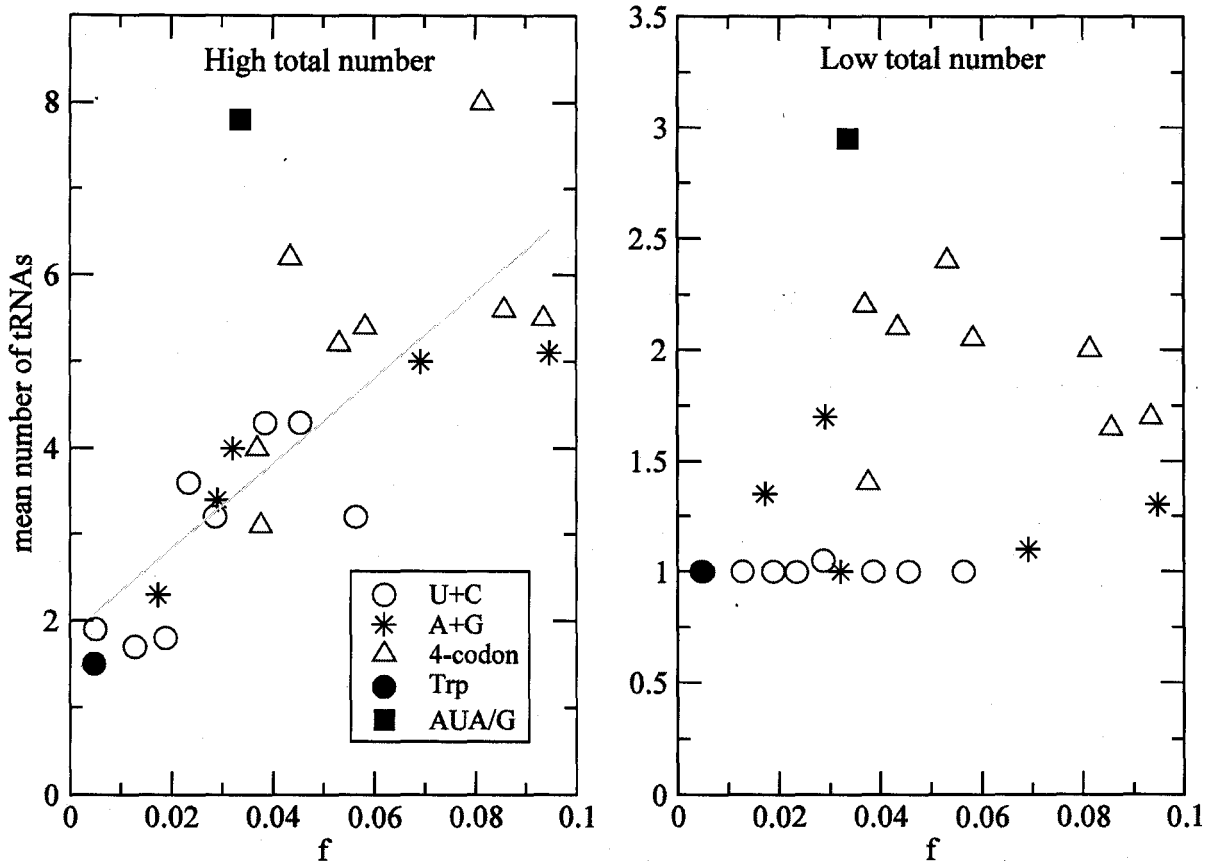


Fig. 3.7. Relationship between the mean number of tRNA genes for each codon family and the frequency of codons in highly expressed genes. High total number genomes are those with 80 or more tRNAs and low total number genomes are those with 40 or fewer tRNAs. Codon families of different classes are labelled by symbols. The line is the linear regression to all points excluding the AUA/G outlier point. There is a clear correlation in large genomes but not small genomes.

3. 6 Variation of selection strength among organisms

The above sections focused on the direction of selection on codon usage, and the way this depends on the tRNA gene copy numbers. We now consider the way the magnitude of selection varies among organisms. The selection parameter S can be directly calculated from the observed codon counts, as in equation 2.13; therefore, it is easy to compare the magnitude of S among organisms, as was done by Sharp *et al.* (2005). However, S is dependent on the combination of tRNA genes. In A+G families and in four-codon families, the sign of S can change. For some tRNA combinations, the rates of two codons could be almost equal, so S will be small, even if there is strong selection acting on this organism to speed up translation. In U+C families, the direction of selection does not usually vary, but S still depends on the number of tRNA gene copies (as in equation 2.21 and 2.22). The copy number varies among genomes in response to the variation in strength of selection, and this has an effect on S. Thus, S is not the best quantity to compare between species.

We would like some measure of the intrinsic strength of selection on translation. As we discussed in section 2.2, the selective advantage or disadvantage of codon i relative to the reference codon is assumed to be proportional to the difference in the times, Δt, between the two codons: $s_i = s_0 \Delta t = s_o(\frac{1}{r_{ref}} - \frac{1}{r_i})$. Here s_o determines the amount by which the fitness changes due to a given saving in translation time, Δt. However, s_0 is not separable from the other unknowns in the theory. It occurs in the combination $K = \frac{2N_e s_o}{c_o k_o}$ as shown in equation 2.17. From equation 2.21, $K = S_C N_G / (1/b_{GU} - 1/b_{GC})$.

S_C and N_G are known. Therefore, the value of K can be estimated for each U+C family and an average can be taken over U+C families, as described in section 3.5. The b parameters are not known, but we assume they are the same for each family; therefore they are just a constant multiplying factor that does not affect the relative values of K in different species.

In Figure 3.6 it was already shown that there is a relationship between K and the doubling time of bacteria. Doubling time was estimated by Rocha (2004) as the minimum doubling time observed for a species under favourable growth conditions. Species with the largest K usually have doubling times of 30 minutes or less, whereas those with the smallest K often have doubling times of more than 10 hours, and sometimes more than 100 hours. It was also shown that the total number of tRNA gene copies in the genome increases from around 30-40 for the species with the smallest K to over 100 for those with the largest K.

Our interpretation of these observations is that K varies among species principally because s_0 varies. The observed growth rates of bacteria differ by several orders of magnitude. As a general rule, rapid growth is presumably an evolutionary advantage; however, the degree to which rapid growth matters seems to vary between species. Slow-growing bacteria survive perfectly happily in some environmental niches, which suggests that other factors are selected more strongly than growth rate in these niches, or growth rate may be simply be limited by scarcity of resources so there is no selection for the ability to grow fast. On the other hand, fast growing species must live in an environment where resources are abundant (at least some of the time) and where rapid growth enables

them to out-compete more slowly growing cells. Protein production rate should be important to the fast growing species. Hence, there is selection to optimize codon usage and to duplicate tRNA genes in fast growing species.

The following crude argument suggests that s_0 should be inversely proportional to the doubling time T. Suppose a synonymous mutation in one codon allows a small time saving Δt. If N_{prot} copies of the protein are translated from this gene in one cell generation, then the total time saved is $N_{prot}\Delta t$. The fitness of the mutant relative to the original should be inversely proportional to the relative doubling times, *i.e.* $1+s = T/(T - N_{prot}\Delta t) \approx 1 + N_{prot}\Delta t / T$. From equation 2.15, we have $s_0 = N_{prot}/T$. Hence we expect K to be proportional to $1/T$ if other factors (N_e, k_o, and c_0) are equal. This argument is clearly oversimplified, but it correctly predicts that translational selection is strongest when the protein expression is highest and when the doubling time is shortest.

One factor that might confound this interpretation is that K is proportional to the effective population size N_e, and this also varies among species. Thus it might be that rapidly multiplying organisms have high K because they happen to have high N_e rather than because they are rapidly multiplying. It is not usually possible to determine N_e as a single parameter, but Lynch and Conery (2003) determined the product $N_e u$ for several species (where u is the mutation rate), including nine of the bacteria in our data set. Fig 3.8 shows K plotted against $N_e u$ for these nine species. There is no correlation (R = -0.037, p = 0.925). Figure 3.8 also shows that there is no correlation between doubling time and $N_e u$ (R = 0.277, p = 0.470), *i.e.* it is not true that rapidly multiplying species have high $N_e u$. In Figure 3.9 We plot K against $1/T$, as suggested by the crude argument above. For

the same nine species (shown as black points), there is a significant correlation ($R = 0.832$, $p = 0.0054$). If all the species are included, the correlation becomes stronger ($R = 0.851$, $p < 0.0001$). This p value should be treated with caution because we have not attempted to correct for the phylogenetic relatedness of the species. The species were initially selected by Sharp *et al.* (2005), to eliminate closely related genomes, and synonymous substitutions occur very rapidly on the timescale of the bacterial evolutionary tree. So these species should be reasonably independent.

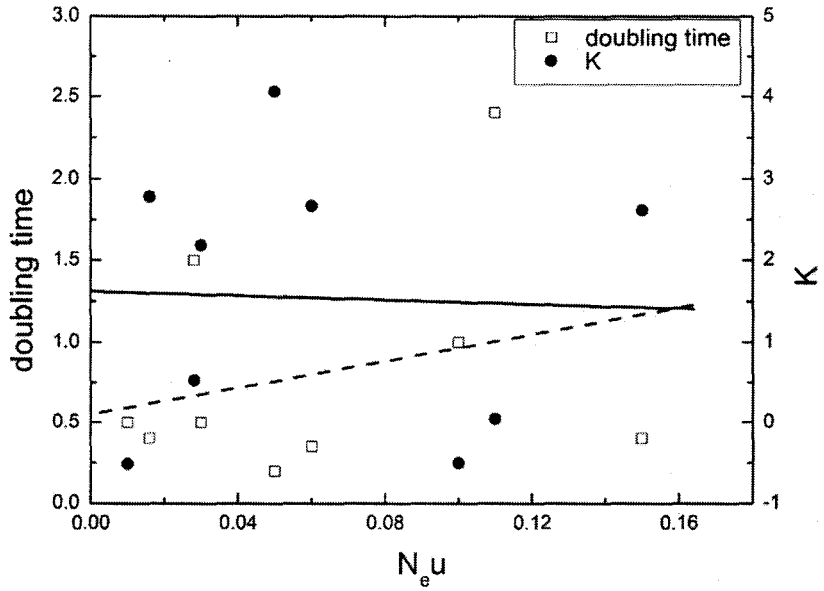


Fig. 3.8. Relationship of doubling time and translational selection strength, K, to effective population size × mutation rate, $N_e u$. Neither quantity is correlated with $N_e u$. The black line is the linear fitting of doubling time vs. $N_e u$, while dash line, of K vs. $N_e u$. No correlations are found for any sets of papameters.

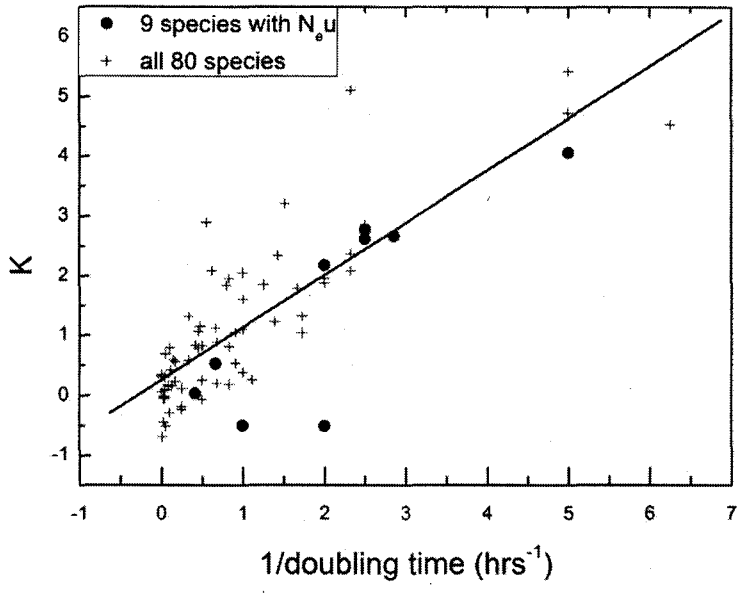


Fig. 3.9. Translational selection strenth, K, is significantly correlated with the inverse of the doubling time (1/T), both for the nine species in which $N_e u$ was measured and for the full data set.

3. 7 Discussion

We have used a simple assumption about kinetics of translation in this thesis, namely that the rate of translation of a codon by a given tRNA is proportional to the tRNA gene copy number and a rate constant that depends on the codon-anticodon combination. Experimental studies of the translation process have shown that there are many steps to the translation of each codon (Rodnina and Wintermeyer, 2001; Blanchard *et al.* 2004; Daviter *et al.* 2006), and complex models of the translation process have been developed

(Heyd and Drew, 2003; Ninio, 2006). There is not yet complete agreement as to how to define the different steps, and most of the measurements focus on differences in rates between cognate and non-cognate codons, whereas it is differences in rates between synonymous codons that are relevant for codon bias. The fact that codon bias occurs in a large number of bacterial genomes means that these rates must indeed differ. Theories like ours can therefore be used to predict which codons should be translated most rapidly, and thus give suggestions for future experiments. In this chapter we have only dealt with two-codon families. But in the same way the SMD theory for four-codon families will be done straightforwardly in the following chapter and a kinetic model can also be specified as a future work.

Here we mention two other recent theories that consider the relative ability of different kinds of anticodon-codon combinations to pair. Dos Reis *et al.* (2004) have developed a tRNA adaptation index to assess the degree to which codon usage in a gene is adapted to the tRNA content of the genome. Xia (2008) has considered coevolution between codon usage and tRNA anticodons in fungi mitochondria. Both papers use parameters that have some relation to our b_{ij} parameters, although they are defined in different ways.

A simplifying assumption made in our theory is that tRNA concentrations are directly proportional to gene copy number. For some organisms, information is available about the concentrations of tRNAs in the cell. It would therefore be possible to use these concentrations explicitly in the theory. We have not done this because it would then be impossible to carry out the statistical survey of large numbers of organisms. In cases

where it has been measured, a rough proportionality between concentration and gene number exists, *e.g.* in *Bacillus subtilis* (Kanaya *et al.* 1999) and *Saccharomyces cerevisiae* (Percudani *et al.* 1997). In *E. coli* more detailed information about regulation of tRNA gene expression is available (Dong *et al.* 1996). When *E. coli* is grown at a variety of different growth rates it is found that the concentration of tRNAs cognate to the most frequent codons increases as growth rate increases, although not dramatically, and the concentrations of tRNAs cognate to less frequent codons remain unchanged with growth rate. This suggests some degree of regulation of tRNA gene expression. One factor causing regulation of tRNA genes is the positioning of genes within the genome. Genes close to the origin of replication may be present in a double dose, whereas those that are further away will be replicated later in the cycle and are less likely to be present in a double dose. This should lead to corresponding variation in the tRNA concentrations. Ardell and Kirsebom (2005) have investigated the dosage effect and also the effect on expression of transcription of tRNAs in operons of several genes.

Although these complex details are interesting, we should not forget that a simple way to regulate the concentration of tRNAs is to duplicate or delete the gene. We presume that duplications and deletions occur randomly with respect to the type of tRNA but selection operates among genome variants with different gene contents. In organisms with strong translational selection, the tRNA gene content is important to the organism and there will be significant selective differences among genome variants with different tRNA copy numbers. Genomes with efficiently coevolved sets of tRNA genes will tend to replace those with less efficient sets. Although there have been many previous studies of

codon usage, ours is the first that gives a theory explicitly describing the co-evolution of codon usage with tRNA gene content and that carries out a large scale survey of the trends in many bacterial genomes that are caused by this co-evolution.

Chapter 4

Anticodon-Codon Interaction and tRNA Modification

Preface

In this chapter, we study the way that codon usage in bacteria depends on the interaction between codon and anticodon. We show that preferred codons do not always correspond to Watson-Crick pairing with the tRNA and that the presence of modified bases at the wobble position can have an important influence on codon usage.

4.1 Introduction

In Chapter 3, we used a simple model of translation kinetics to calculate mean translation times of each codon, and proposed that the selection coefficient for one synonymous codon over another was proportional to the difference in these times. This theory predicts which codon is preferred for each combination of tRNA gene copy numbers. Codon frequencies and tRNA copy numbers evolve towards co-adapted stable states in which neither can change without the other. We showed that, in many cases, there is more than one possible stable state. This explains why the direction of selection sometimes differs between codon families of the same type in the same organism (*e.g.* for A+G families, the A codon might be preferred for one amino acid but the G codon for another). We also included the possibility of tRNA gene duplication and deletion, and showed that genes will be duplicated for organisms in which the intrinsic strength of selection is high. Our theory therefore explains why high tRNA gene copy numbers are found in species with

strong codon bias, and why both high copy numbers and strongly biased codon usage are found in fast-multiplying organisms.

The key unknown parameters in our theory are the relative rates of pairing of the different anticodon-codon combinations, *i.e.* the b parameters in Figure 1.2. The main objective in this chapter is to deduce as much as possible about these rate parameters by comparing the observed codon usage to the theory. We show that the presence of modified bases in the anticodon influences the rate parameters and leads to observable effects on codon usage.

4.2 Analysis of codon usage in U+C codon families

In previous Chapter, we already considered U+C codon families. We summarize this here, as it is necessary to compare this case with more complex cases considered later. A set of 80 genomes was used that spans the full range of complete bacterial genomes, as previously selected by Sharp *et al.* (2005). We obtained S_C for the C codon in each family, using the U codon as reference. We found that there is consistent preference for the C codon in most species and the selection parameter is given by equation 2. 21. Preference for C indicates that $b_{GC} > b_{GU}$. Our interpretation is that the strongly interacting GC pair is processed more rapidly by the ribosome than the weakly interacting GU pair.

In some cases in table 3.3, the U codon is found to be more frequent in the high expression genes; hence S_C is found to be negative in Equation 2.21. This could be explained if b_{GC} were less than b_{GU} for some tRNAs. However, we do not think this is true. The cases with negative S_C occur in species with slow growth rates, few tRNAs, and

generally weak translational selection. Thus we interpret the increase of U in the high expression genes in these cases as being a mutational effect that is not accounted for in the model, *e.g.* base frequencies may differ between leading and lagging strands or according to position of a gene relative to the origin of replication, and high expression genes may be non-randomly positioned on the genome. These effects might be present to some extent in all species, but they would show up as anomalies in cases where translational selection is weak. As described by Chapter 3, 14 of the original 80 species were excluded from the subsequent analysis because they did not show a consistent preference for C in U+C families, and it appeared that translational selection was weak in these species. The remaining 66 species showed a consistent preference for C, and it was assumed that translational selection was a significant effect.

4.3 Analysis of codon usage in A+G codon families

The following analysis was carried out with the 66 species in which consistent evidence for translational selection was identified in the U+C families. We included the A+G codon families for Leu(UUR), Gln(CAR), Lys(AAR), Glu(GAR) and Arg(AGR). For each amino acid in each genome, the observed codon usages π_i and ϕ_i in the low and high expression genes were obtained from sequence data (using Equations 2.11 and 2.12). We used A as the reference codon, and determined the selection strength of G relative to A using Equation 2.13. We define the preferred codon as the one whose frequency increases in the high expression genes. This is the one with the higher selection coefficient, *i.e.* G is preferred if S_G is positive and A is preferred if S_G is negative. Cases

were grouped according to the combination N_U:N_C of tRNA genes that is present for that amino acid. Table 4.1 shows the number of cases for which A and G codons are preferred for each combination of tRNAs. The mean values of the codon frequencies in high and low expression genes are also shown for each tRNA combination. Frequencies that increase in the high expression genes are underlined.

The selection strength of G codon relative to A is given by equation 2.13, which may be positive or negative, depending on the number of tRNA gene copies of the two types, and on the values of b parameters. In Chapter 3 it is showed that the direction of selection in the majority of cases could be explained if we set $b_{UA} = b_{CG} = 1$, and $b_{UG} = 0.4$. This is a plausible guess if we assume that wobble pairing is weaker and slower than Watson-Crick pairing. In this chapter, we want to use the data on codon usage to derive information on the relative rates, rather than assuming some particular values at the outset.

In Table 4.1, we see that when $N_U > N_C$, the A codon is preferred in the majority of cases and the mean frequency of A is higher in high expression genes than in low expression genes. In contrast, when $N_U \leq N_C$, the G codon is preferred in the majority of cases and the mean frequency of G is higher in high than low expression genes. This can be explained using equations 2.19 and 2.22. The most important cases that give information on the b parameters are N_U:N_C = 1:1 amd 2:1. As G is preferred in the 1:1 case, we obtain $b_{UG} + b_{CG} > b_{UA}$, and as A is preferred in the 2:1 case, we obtain $2b_{UG} + b_{CG} < 2b_{UA}$. These inequalities can be rewritten as $(1\text{-}b_{UG}/b_{UA}) < b_{CG}/b_{UA} < 2(1\text{-}b_{UG}/b_{UA})$, which defines the region between the two solid lines in Figure 4.1. Most of the other combinations (such as 3:1 and 1:2) give inequalities that are less restrictive than the two

above. Therefore, any set of parameters that falls between these lines will also correctly explain the direction of selection for the other tRNA combinations. Note that the parameters that we chose as examples in Chapter 3 fall in the middle of this range – $b_{CG}/b_{UA} = 1$ and $b_{UG}/b_{UA} = 0.4$. The only case that is more restrictive than these two is 3:2, for which A is preferred. This gives $b_{CG}/b_{UA} < 3/2(1-b_{UG}/b_{UA})$ which restricts parameters to the region below the dotted line in Figure 4.1. However, there is just one such example of the 3:2 case, so this condition is not as well substantiated as the other two.

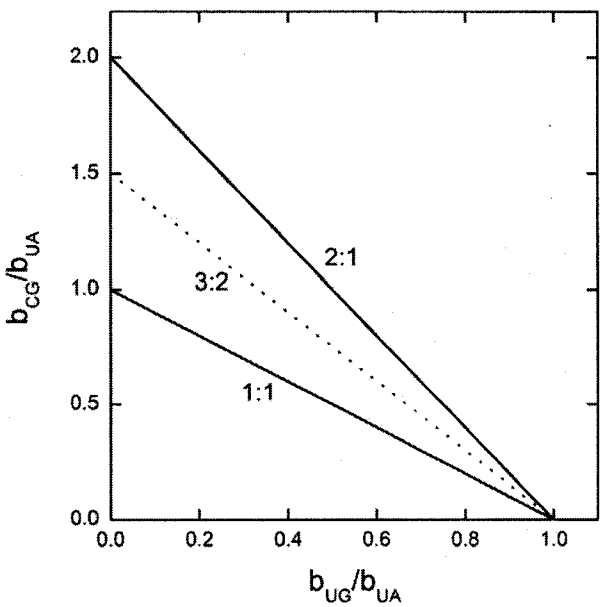


Fig 4.1. Regions of parameter space for the relative rates of translation in A+G codon families.

Three cases at the bottom of Table 4.1 have $N_U = 0$. This is impossible according to our model if only the U tRNA can translate the A codon. The two cases with $N_U:N_C =$

0:1 do not have particularly strong codon bias, so it may be that there is a mis-annotation of the tRNA genes in these cases, and there is really a wobble-U tRNA present. The single case with $N_U{:}N_C$ = 0:2 stands out because the A codon is completely absent from the high expression genes, and is very rare in the low expression genes. It is therefore possible that this case is real. The low value of π_A would then indicate selection against the A codon that was strong enough to be observed in the low expression genes rather than just the high expression genes.

Combining all tRNA combinations, the theory predicts the correct direction of selection in 238 out of 330 cases (72%) when parameters in the appropriate range are used. Many of the cases that do not follow the prediction occur for the combinations 1:1 and 2:1. The tipping point in the direction of selection occurs between the 2:1 and 1:1 cases, so we expect that the rates of the two codons are close to one another for these tRNA concentrations. The predicted direction of selection will therefore be sensitive to the assumption that the tRNA concentrations are proportional to the number of gene copies. Small differences from this proportionality might tip selection in the other direction. It should also be remembered that the relative pairing rates might depend on features of the tRNA other than the wobble base, so there may not be one single set of b parameters that is exact for every tRNA with the same wobble base. There are also a fairly large number of examples that do not follow the majority prediction in the 1:0 case. This case occurs where selection is weak, because duplicate tRNAs would be favoured if selection were stronger (as discussed in Chapter 3). Hence, the larger number of exceptions in this case is probably because of mutational effects that are not accounted for

in the model that show up only when selection is weak (as described in the section on U+C families).

The genomes considered in this analysis vary in GC content considerably. Table 4.1 shows that there is a correlation between GC content and tRNA gene copy numbers. The tRNA combinations where A is preferred correspond to cases where $\pi_A > \pi_G$ in low expression genes, whereas the reverse is true for combinations where G is preferred. This means that the selection arising from the tRNAs is acting in the same direction as the mutational bias in the majority of cases. Hence, the mean frequencies in high expression genes are more strongly biased in the same direction as in low expression genes. The following argument suggests why this should be the case. Consider a slowly-multiplying organism with weak translational selection. We expect this to have only one tRNA for a given codon family, and we know that this must be a wobble U tRNA whatever the GC content of the organism, *i.e.* we expect $N_U{:}N_C$ = 1:0. Now suppose this organism moves into an environment where rapid multiplication is required and translational selection is stronger. It then pays to have a second tRNA because the benefit from increased translation rate outweighs the cost of the additional tRNA (as shown in detail in Chapter 3). The second tRNA could be a simple duplication to give the 2:0 combination, or a duplication followed by an anticodon mutation in one of the copies, resulting in the 1:1 combination. If there is a pre-existing mutational bias in the codon frequencies of the low expression genes, we would expect the additional tRNA to follow this existing bias because the greatest speed-up in translation will occur if the additional tRNA matches the codons that are already most frequent. Thus, when $\pi_A > \pi_G$, we expect to see tRNA

combinations with increased N_U, and when $\pi_G > \pi_A$ we expect to see increased N_C. Table 4.1 shows that the average behaviour does indeed follow these expectations.

However, there are a significant number of cases that do not follow the average trend. Of the 330 cases considered, the preferred codon is less frequent in the low expression genes in 92 cases (28%). We have previously shown in Chapter 3 that multiple stable states of codon usage and tRNA copy number can exist because the two quantities are co-adapted. In a stable state, neither tRNAs nor codon frequencies can change without decreasing translational efficiency. Some of these stable states are paradoxical, in the sense that the direction of selection is opposite to that of mutation bias. The observation of these paradoxical states in the data is an important confirmation of the coevolution theory in our previous chapter. The most likely way that a paradoxical state could be reached is that a genome was initially in an state where the selection followed the mutation bias, but the direction of mutation bias subsequently changed because of changes in the enzymes responsible for DNA replication or in the internal biochemistry of the organism. The codon frequencies in low expression genes would then change to follow the new mutation bias, but those in high expression genes would remain biased in the original direction because of selection from the existing tRNAs. Only if the mutational bias in the opposite direction became very strong would the existing coevolved state of tRNA copy numbers and codon usage become unstable (see Chapter 3 for conditions on stability). Coevolution of tRNAs and codon usage can therefore exhibit hysteresis.

4.4 Analysis of codon usage in four-codon families

We now consider four-codon families in the same set of 66 bacterial genomes. Eight codon families are considered: Leu (CUN), Val (GUN), Ser (UCN), Pro (CCN), Thr (ACN), Ala (GCN), Arg (CGN) and Gly (GGN). Although Leu, Ser and Arg have six codons, these can be divided into a group of two and a group of four that are translated by separate tRNAs. Therefore, we consider the group of four codons for these amino acids in this section and ignore the group of two. The pattern of base pairing expected in four-codon families is shown in Figure 2.1c, and the relative rates of translation are as in Equation 2.20. Cases in which a wobble-A (or I) tRNA occurs are excluded, but will be considered as a special case at the end of this section.

We choose U as a reference codon because U often has a high frequency in the high expression genes, and measure the selection coefficient of the other codons with respect to this, as in equation 2.13. We define the preferred codon to be the one with the highest S. As before, the preferred codon is not necessarily the most frequent codon because mutation bias and selection do not necessarily act in the same direction. The reference codon has $S_{ref} = 0$. It may be that the reference codon is preferred, in which case the other three codons all have $S < 0$.

In Tables 2-4, each line corresponds to one combination of tRNA gene copy numbers. Table 4.2 includes cases where only N_U is non-zero. Table 4.3 includes cases where N_G and N_U are non-zero. Table 4.4 includes cases where N_C is non-zero together with varying combinations of N_G and N_U. The fact that there are many cases where only N_U is non-zero (Table 4.2) clearly shows that wobble-U tRNAs can pair with all four

codons. In this table, the A and the U codons are most frequently preferred. Based on Watson-Crick pairing, we would expect A to be preferred. It is surprising that there are almost as many cases when U is preferred as when A is preferred. In a four-codon family, the number of codons that increase in frequency in the high expression genes can be one, two or three. Counting the preferred codons shows which codon increases the most, but does not tell us whether other codons also increase. The average codon frequencies in Table 4.2 show that, typically, both U and A codons increase (as shown by the underlined figures in the table), whereas G and C decrease. We also find that the S value for the C codon is usually the lowest, *i.e.* C usually decreases the most in the high expression genes.

In terms of the rate parameters, these results show that b_{UA} and b_{UU} are both high and are roughly equal to one another. Both of these are higher than b_{UG}, and all three are higher than b_{UC}. We tried in several ways to obtain more quantitative estimates of these rate parameters, including by directly fitting the observed codon frequencies to a maximum likelihood model that uses the b_{ij} parameters. It proved to be difficult to estimate one average set of parameters that applies across many cases because there is considerable scatter in observed codon bias even among cases with the same tRNA combination. This could arise from variations in tRNA concentrations that are not directly proportional to copy number, or because the relative pairing rates are different for tRNAs with the same wobble base because of other structural differences. Therefore, at this point, we work on the qualitative result that is typical for wobble-U tRNAs in four codon families - $b_{UA} \approx b_{UU} > b_{UG} > b_{UC}$. The high value of b_{UU} is an unexpected prediction from our analysis, which could be tested in future experiments on translational kinetics.

Although a high rate of UU pairing is unexpected, there is no intrinsic problem with this if it occurs in a four-codon family. However, b_{UU} cannot be large in codon boxes that are split between two amino acids. For example, a wobble-U tRNA for the Gln A+G family cannot have a high rate of pairing with the U codon for His. Therefore, these results show that wobble-U tRNAs in A+G families behave differently from those in four-codon families. The main reason for this is that the U base is modified in different ways in the two cases. Modified bases will be discussed in detail in a later section.

Another point worth noting in Table 4.2 is that the situation where only wobble-U tRNAs are present occurs when π_U and π_A are high in the low expression genes. This makes sense because if both b_{UA} and b_{UU} are high, then a single tRNA type is good enough to translate the codons that are most frequent. Selection will then act to increase the frequency of U and A in the high expression genes. Therefore, the direction of translational selection is in the same direction as the mutational bias for most of the cases in Group 1.

The examples in Table 4.3 are listed in order of increasing ratio of N_G:N_U. In the upper half of the table, $N_G < N_U$. Here, both U and A codons increase in frequency on average in the high expression genes, as with Table 4.2. U and A are most frequently the preferred codons, although there are now substantially more cases of U than A. In the lower half of Table 4.3, $N_G \geq N_U$. Here, only the U codon increases in the high expression genes and there is a large majority of cases where U is the preferred codon. Moving down Table 4.3 shows the effect of increasing N_G relative to N_U. We expect that wobble-G tRNAs pair principally with U and C codons. These results give information about the

relative sizes of b_{GC} and b_{GU}. We already saw that in two-codon U+C families, there is a consistent preference for the C codon, from which we concluded b_{GC} is substantially greater than b_{GU}. If b_{GC} is also high in four-codon families, we might expect to see the C codon being preferred in cases where N_G is large. However the U codon is preferred in all the cases in the bottom half of Table 4.3, including those where N_G is very large. The most extreme ratio of N_G:N_U is 6:1. The fact that $\rho_U > \rho_C$ in this case tells us $6b_{GU} + b_{UU} > 6b_{GC} + b_{UC}$. This can be rearranged as $b_{GC} < b_{GU} + (b_{UU} - b_{UC})/6$, *i.e.* b_{GC} cannot be too much larger than b_{GU}. In fact, from the Group 3 results, it is even possible that $b_{GC} < b_{GU}$ in four-codon families. Unfortunately, we cannot tell this, because there is no case where wobble-G tRNAs occur without wobble-U tRNAs. Thus, we cannot observe the codon preference that would arise from the influence of the wobble-G alone. These results suggest the relative rates of pairing of wobble-G tRNAs with U and C codons may be different in two-codon and four-codon families, and that the preference for the C codon may be weaker in four-codon families.

Wobble-G tRNAs can also pair with A codons in some circumstances (see the section on modified bases). Therefore, the possibility of GA pairing is allowed in Figure 2.1c and Equation 2.20. However, Table 4.2 shows that b_{GA} must be small. The principal effect of increasing N_G is to reduce the preference for A and to increase the preference for U. This is what we would expect if b_{GA} is small. If GA pairs are possible, it should be possible for a combination of wobble-G and wobble-C tRNAs to translate all four codons if there are no wobble-U tRNAs. However, there are no observed cases where $N_G > 0$, N_C

> 0 and $N_U = 0$. The probable reason for this is that the wobble-G tRNA on its own is not sufficient to translate the A codon. Once again, this shows b_{GA} must be small.

The G and C frequencies in the low expression genes in Table 4.3 are slightly higher than those in Table 4.2. This shows that if the mutational process is such that the number of C codons is moderately large, then it pays to have a wobble-G tRNA to translate them. As discussed above, C is the least efficiently translated codon when only wobble-U tRNAs are present. Nevertheless, the presence of the wobble-G tRNA does not automatically create selection for the C codon, because b_{GC} does not seem to be particularly large compared to other rates, as discussed above.

In Table 4.4, the cases where $N_C = 1$ are rather similar to Tables 2 and 3. The preferred codon is usually U. If $N_G = 0$ or $N_G < N_U$ then both U and A codons increase in frequency in the high expression genes, whereas when $N_G > N_U$, usually only the U codon increases in frequency in the high expression genes. The 1:1:1 case is the most frequent combination in this table. The preferred codon is usually U but is not very consistent. In this case, the rates of the four codons are probably not that much different from one another and the results are influenced by variations in tRNA concentrations that are not strictly proportional to gene copy number.

Cases where $N_C \geq 2$ (at the bottom of Table 4.4) are fairly rare. When there is a high proportion of wobble-C tRNAs, such as the cases 1:1:2, 1:1:4 and 1:2:2, the G codon is often preferred. This suggests that b_{CG} is reasonably large. However, in cases where both N_G and N_C are in high proportion, such as 2:1:2, 3:1:3 and 3:1:2, the U codon is preferred over G. So this tells us that b_{CG} cannot be particularly large in comparison to

b_{GU} and b_{UU}. Cases where all three tRNAs exist usually correspond to GC-rich codons in the low-expression genes. This is the pattern we would expect, because wobble-G and wobble-C tRNAs are more useful for translation in high GC species. However, since the U codon is preferred in many of these cases, these are examples where selection is acting in the opposite direction to the mutation bias.

There are several cases where a strong GC skew is apparent in the low expression genes (*i.e.* the G and C frequencies are widely different). For example, π_C is much greater than π_G in cases where N_G is in high proportion (such as 6:1, 6:2, 7:2 and 11:2 in Table 4.3), whereas π_G is much greater than π_C in cases where N_C is in high proportion (such as 1:1:2 and 1:1:4 in Table 4.4). This shows that tRNA duplications are influenced by existing mutational bias. We presume that the reason the GC skew exists in the low expression genes is because of context-dependent mutation. This means that the π frequencies can be different in different four codon families. We will not pursue this point here, although we have discussed it in detail in the case of codon usage in mitochondrial genes (Jia and Higgs, 2008).

Table 4.5 shows the cases of four codon families in which there are tRNAs with an A at the wobble position in the tRNA gene sequence. The A base is known to be modified to inosine (I) in many cases in the tRNA molecule, but we will use the notation N_A for the number of copies of this gene. Almost all of these cases occur for the Arg AGN codon family, but there are also a small number of examples for Leu CUN and one single example for Thr ACN. There is a single example where N_A:N_G:N_U:N_C = 1:0:0:0, which shows that the I can pair with all four codons. More typically, this tRNA occurs with a

wobble C tRNA, which suggests that the I pairs efficiently with U, C and A codons, but not with G codons, and that the wobble C tRNA is usually required to translate the G codon. In almost all the examples in Table 4.5, the preferred codon is U. Thus, b_{IU} must be the largest of the rates involving the wobble-I base. It is also seen that U is preferred relative to G, even though there is usually a wobble-C tRNA present to translate the G codon. Thus, b_{IU} must be high with respect to b_{CG} as well.

4.5 The role of modified bases at the wobble position

There are many cases where the wobble base of the tRNA is modified in a way that influences anticodon-codon pairing (Curran, 1998; Agris 2004, 2008). Although a considerable amount is known about the function of these modifications in some species, we do not have systematic information about which modifications occur for all the genomes analyzed in this chapter. We therefore classified the tRNAs according to the base at the wobble position before any modifications, as this is directly observable in the tRNA gene sequence. In this section we consider the way that modified bases might influence the interpretation of the codon preference data above.

A general mechanism that applies to several types of modifications (Agris 2008) is that the modification reduces the conformational flexibility of the anticodon stem-loop. When the tRNA interacts with the codon at the A site in the ribosome, it needs to be recognized as a correctly matching tRNA. Recognition requires a reduction in entropy of the anticodon as it fits into place. By reducing the entropy of the anticodon loop in its unbound state, the modification lowers the entropic barrier for recognition; hence the

recognition step is speeded up. The magnitude of this effect can depend on the base at the third codon position; therefore a modification can change the relative rates of translation of alternative synonymous codons. Modifications can also act to prevent pairing of anticodon-codon combinations that would occur with the unmodified base. This is important in codon boxes that are split between two amino acids because it prevents mistranslation of a codon by a tRNA for another amino acid. These effects are very specific; thus it is necessary to consider each case separately.

In U+C families, the G base at the wobble position is modified to queuosine, Q, in tRNAs for Tyr, His, Asn and Asp. This is found in all domains of life (Romier *et al.* 1998), although the enzymes responsible may not be the same in all domains (Morris and Elliott, 2001). We therefore presume that this modification occurs in the majority of the bacteria in our sample. However, a detailed study of *Mycoplasma capricolum* (Andachi *et al.* 1989) shows that the G remains unmodified, and other cases outside bacteria are also known where the G in these tRNAs is unmodified (Morris and Elliott, 2001; Jühling *et al.* 2009). Furthermore, the G is unmodified in tRNAs for other U+C codon families (Phe, Cys, and SerAGY) and for the three-codon Ile family. Our results above (and also those of Sharp *et al.* 2005 and Chapter 3) show that the C codon is preferred in U+C families both in the cases where the Q modification is common and where it does not occur. Urbonavicius *et al.* (2001) studied translation in *E. coli* using tRNAs with and without the Q modification. They concluded that both Q and G interact faster with the C codon than the U codon, which agrees with our observations from codon frequencies. However, Meier *et al.* (1985) found that $tRNA^{His}$ in *Drosophila* showed a definite preference for the

C codon when the G was unmodified, but a slight preference for U when the Q modification was present. Also Morris *et al.* (1999) showed by molecular modelling that the Q modification promotes the ability of this type of tRNA to pair with the U codon. It seems likely that the reason for the presence of the Q modification is to increase the efficiency of pairing with the U codon, *i.e.* the Q modification probably decreases the difference in rates between the C and U codons. However, our observation is that the C codon is still preferred, even when Q is present; therefore we conclude that the Q modification does not change the direction of the codon preference, at least in bacteria, so it does not make much difference to the interpretation of our codon usage data.

The Q modification is relevant with respect to the origin of variant genetic codes in mitochondrial genomes (Yokobori *et al.* 2001; Sengupta *et al.* 2007). In some animal mitochondria, tRNAs with an unmodified G translate A codons in addition to U and C. This occurs for the AUA Ile codon in some species, although in many other species AUA is reassigned to Met. The Unassigned Codon mechanism that we have proposed to explain the reassignment of AUA (Sengupta and Higgs, 2005; Sengupta *et al.* 2007) relies on the fact that the wobble G tRNA can pair with the A codon to some extent. There are some mitochondrial codes in which AGR has been reassigned from Arg to Ser, and wobble-G tRNAs seem to translate AGA codons as Ser in these cases. The Q modification is found in many mitochondrial tRNAs for Tyr, His, Asn and Asp, just as in bacteria. Thus it may be that one of the functions of the Q is to restrict the mispairing of wobble-G tRNAs with the A codon in codon boxes that are split between two amino acids. Another variant mitochondrial genetic code supports this point. In some species, the AAA

codon is reassigned from Lys to Asn. This is associated with the loss of the Q modification from the $tRNA^{Asn}$, which enables the unmodified G to interact with the A codon. The examples mentioned here indicate the GA pairing is possible in some cases, which is why we included this as a possibility in our kinetic model (Figure 2.1). As we discussed in the results section, however, the codon usage data show that the b_{GA} rate is small in bacteria compared with b_{UA}, and we did not find examples in bacteria where a wobble-G tRNA is the only tRNA that translates an A codon.

One of the surprising results from the codon usage analysis is that the U codon is often preferred in four codon families when wobble-U tRNAs are the only ones (Table 4.2), or the most frequent ones (Table 4.3). This shows that there is an unexpectedly high efficiency of pairing of the UU combination, with a rate b_{UU} that is similar to b_{UA} and higher than b_{UG}. Nevertheless, when wobble-U tRNAs occur in two-codon A+G families, the UU pairing rate cannot be high, because this would lead to high rates of mistranslation. Thus, the codon usage data show that wobble-U tRNAs behave differently in two- and four-codon families. Base modifications seem to be an important part of the explanation for this. In tRNAs for four-codon families, the U base at the wobble position is modified to 5-methoxyuridine (mo^5U) or uridine-5-oxyacetic acid (cmo^5U) in most bacteria (Curran, 1998; Agris, 2004, 2008; Jühling *et al.*, 2009). We will denote this class of mutations as xo^5U. In tRNAs for two-codon families, the U base usually has a 5-methylaminomethyl modification and often a 2-thio modification as well. Examples occurring in bacteria are 5-methylaminomethyl uridine (mnm^5U), 5-carboxymethylaminomethyl uridine ($cmnm^5U$), 5-methylaminomethyl 2-thiouridine

(mnm^5s^2U) and 5-carboxymethylaminomethyl 2-thiouridine ($cmnm^5s^2U$). We denote this class as xm^5U. The two classes of modified U bases function in different ways, as the following examples show.

Several cases of xo^5U modifications have been studied experimentally. Lustig *et al.* (1993) showed that $tRNA^{Gly}$ would only bind with G and A codons when the U was unmodified, but with all four codons when the modification was present. Kothe and Rodnina (2007) found that for $tRNA^{Ala,}$ the modification was essential to allow binding to the C codon. Näsvall *et al.* (2007) found that xo^5U tRNAs for Pro, Ala and Val were able to read all four codons, although the similar tRNA for Thr was unable to read the C codon. The presence of xo^5U also improved the reading of the G codon. In all the above cases, the function of the modification is to expand the ability of the tRNA to translate all four codons. Similar results are also found by Weixlbaumer *et al.* (2007) and Vendeix *et al.* (2008). Nevertheless, in *Mycoplasma* the wobble-U base is unmodified in tRNAs for four-codon families (Andachi *et al.* 1989), and the same is also true in animal mitochondria (Yokobori *et al.* 2001, Jia and Higgs, 2007). This shows that an unmodified U can pair with all four codons, at least to some extent. These examples show that there are details of the behaviour of unmodified U tRNAs that seem to depend on which particular tRNA is considered. Lustig *et al.* (1993) also found that U can pair with all four codons in some structural contexts but not others. Nevertheless, the main conclusion from these examples is that the xo^5U modification enhances the ability of the tRNA to pair with all four codons.

In contrast, experiments show that the xm^5U modification restricts pairing to only A and G codons, as is necessary to prevent mistranslation in two-codon families. Experiments on $tRNA^{Lys}$ in E coli (Hagervall *et al.* 1998) showed that the xm^5U modification reduces the rate at which this tRNA misreads Asn codons. However, Ashraf *et al.* (1999) and Yarian *et al.* (2002) found that the 5-methylaminomethyl and 2-thio modifications were essential to allow binding of the $tRNA^{Lys}$ to its own A and G codons, so these modifications seem to help with correct codon pairing as well as in elimination of mispairing. For $tRNA^{Glu}$, these modifications also alter the relative affinities of pairing to own A and G codons (Krüger *et al.* 1998). We are not aware of any examples of wobble-U tRNAs for A+G families in bacteria that lack the xm^5U modification. It appears to be significant that the xm^5U modifications occur even in *Mycoplasma* (Andachi *et al.* 1989), whereas the xo^5U and Q modifications discussed above have both been lost, presumably due to deletion of the genes for the modification enzymes when the reduction in genome size occurred in this lineage. If we assume that an unmodified U can pair with all four codons to some extent, this explains why the xo^5U modification is not essential and can be lost, and also explains why the xm^5U modification is essential and cannot be lost, because otherwise mispairing would occur.

One reason why the two types of modification function in different ways is because they have different effects on the three dimensional configurations of the ribose. The ribose ring is non-planar, and exists in C3'-endo and C2'-endo configurations. In the C3'-endo configuration, Watson-Crick UA pair, and the 'normal' wobble pair UG can form, whereas the non-standard UU pair can only form when the ribose is in the C2'-endo

configuration. Yokoyama *et al.* (1985) showed that the xo^{5}U modification increases the stability of the C2'-endo form relative to the C3'-endo form; hence it makes UU pairing easier. On the other hand, the xm^{5}U modification makes the C2'-endo form less stable; hence it prevents UU pairing. This is consistent with our observation from the codon usage data that UU pairing is fast and U codons are often preferred when the xo^{5}U modification is present. Yokoyama *et al.* (1985) did not give a structure for the UC pair, but we know that this must occur in four codon families as there are many cases in our data where this is the only tRNA (Table 4.2). Experiments also demonstrate formation of the UC pair (Näsvall *et al.* 2007; Kothe and Rodnina, 2007). Table 4.2 shows that the C codon is the least preferred in cases where only wobble-U tRNAs are present, *i.e.* it decreases in frequency in the high expression genes relative to all the other three codons. Therefore, we conclude that UC pairing is possible but is still weak, even in the presence of the xo^{5}U modification.

It will be seen from the results tables that wobble-C tRNAs are much less frequent than wobble-U and G tRNAs. It is presumed that C only pairs with G codons. This is essential for Met and Trp, which have only one codon, but in two- and four-codon families, wobble-C tRNAs are usually an optional extra, because the wobble-U tRNA can pair fairly well with the G codon. C bases are usually unmodified at the wobble position. An interesting exception is 5-formyl cytidine (f^{5}C), which occurs in tRNA^{Met} in some animal mitochondria, and permits this tRNA to interact with both AUA and AUG in species where AUA is reassigned to Met (Yokobori, 2001; Sengupta *et al.* 2007).

In Table 4.5, N_A denotes the number of genes with A at the wobble position, but the A is almost always modified to I in the mature tRNA (Grosjean *et al.* 2010). I is usually thought to pair with U and C and to a lesser extent with A, but not with G (Curran, 1998). For this reason the wobble-I tRNAs usually occur in combination with wobble-C (or sometimes wobble-U) tRNAs that translate the G codon (Table 4.5). The U codon is preferred in almost all these cases. Also the A codon decreases in frequency in high expression genes, which is consistent with there being weak interaction between I and A. It is not clear why an unmodified A is rare at the wobble position. Boren *et al.* (1993) showed that a $tRNA^{Gly}$ with a wobble position that was changed to A was able to read all four codons. They speculated that wobble-A bases are generally avoided because A would be indiscriminate. This argument makes sense in split codon boxes but does not apply in four codon families. Also, if A pairs well with four codons, there must be a reason why it is necessary to modify it to I. Presumably, the I modification must speed up translation of at least the U and C codons relative to the unmodified A, but the reason for avoidance of the unmodified A still seems rather unclear.

All these modifications are summarized in Table 4. 6 for different bases those are found so far. There are not many experiments on base modifications. However, this table already shows that the types of modification are very diversified but can be grouped into just one or two category for each base. The function of modification is to enhance the translation speed or accuracy or both. In Fig. 4.2, the three patterns are put together: The

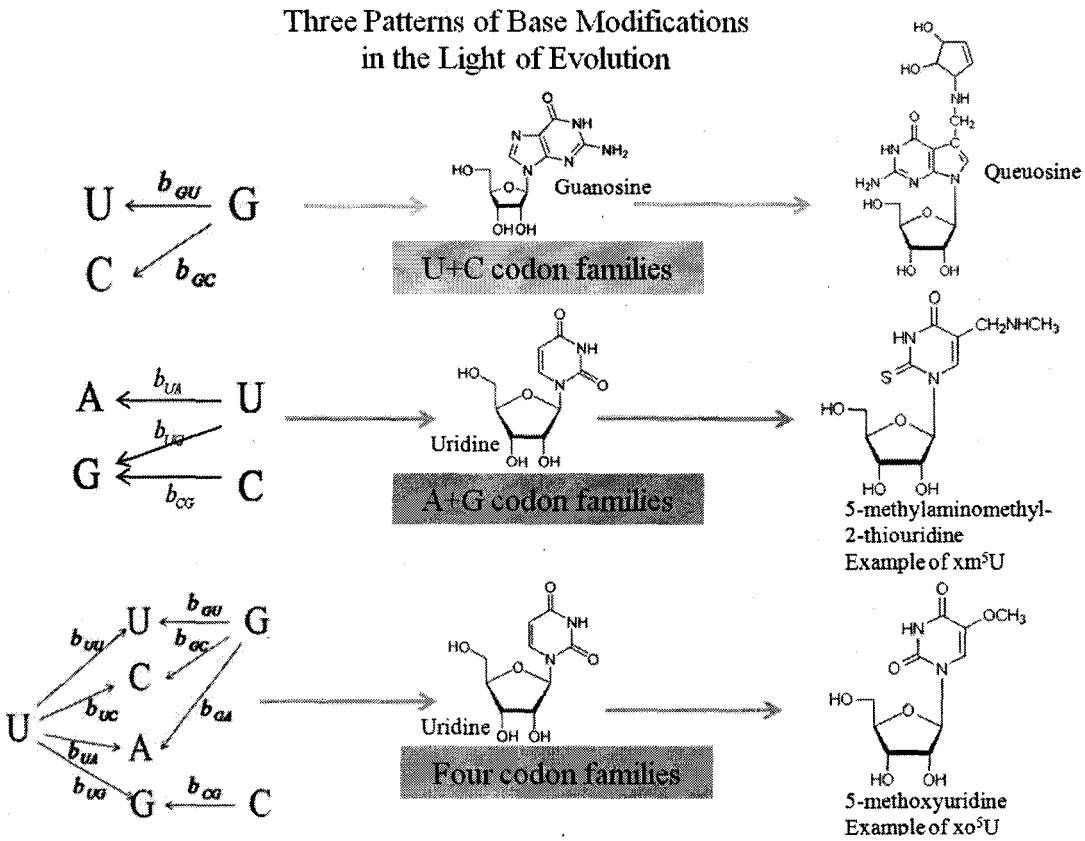


Fig. 4.2 Three patterns of base modifications in the light of evolution: the three translation kinetic patterns on the left, the three codon family patterns in the middle and the three groups of base modification patterns (They are Q, xm^5U and xo^5U. Here only some typical modifications are shown) on the right. During evolution these three groups of specific modifications are introduced in different codon families which follow different translation patterns to make translation fast and/or accurate.

patterns of tRNA modification, the patterns of three codon families and the patterns of translation. The interesting point is that they are related to each other respectively, that's to say, different tRNA modifications are introduced for different codon families which follow different translation patterns to make the translation process fast or/and accurate

during evolution. This is the central point of this chapter and it explains the role of tRNA in the light of evolution.

We have focused on modifications at the wobble position (tRNA position 34) because these have a direct interaction with the third codon position. Base modifications in other positions may also have an essential effect on translation speed and/or accuracy, as have been introduced in section 1.8.

Finally in this section, we note that tRNAs with the same wobble position base may differ in their ability to pair with alternative codons because of structural differences that have nothing to do with base modifications. Positions 36 and 35 of the tRNA pair with positions 1 and 2 in the codon. These are usually standard Watson-Crick pairs with unmodified bases. The strength of these interactions depends on which base pairs are involved. Lehmann and Libchaber (2008) emphasized the importance of an intra-molecular interaction between positions 35 and 33 in the anticodon. They distinguished between strongly and weakly interacting codon boxes based on the strength of the interactions at the 36-1, 35-2 and 35-33 positions. They argued that a single tRNA can pair with four codons in a strongly interacting case, but not in a weakly interacting one. For this reason weakly interacting codon boxes can be split between two amino acids, whereas strongly interacting boxes must remain as four-codon families. This argument is interesting in the context of the evolution of the genetic code (Lehmann and Libchaber, 2008; Higgs, 2009), but it is also relevant in the current discussion of codon usage because it provides another explanation of why the wobble-U tRNAs can translate all four codons in four-codon families, but not in two-codon families.

4.6 Conclusion

In conclusion, the idea of relating codon usage to tRNA concentrations dates back to some of the earliest papers that detected codon usage bias. However, very few studies have considered the nature of the anticodon-codon interaction and the modified bases in the anticodon. Previous papers have tended to assume a one-to-one relationship between codons and tRNAs. Our theory is unique in that it builds on the essential feature that each tRNA interacts with more than one codon, and that often there is more than one tRNA that interacts with the same codon. This has allowed us to relate observations of codon usage more closely to tRNA structure and function and to experimental measurements on translation kinetics and the effects of modified bases on translation.

It is also a rather old idea that the codons that are used frequently in highly expressed genes such as ribosomal proteins are the ones that are preferred by translational selection. This is incorporated into measures of codon bias such as Codon Adaptation Index. However, our theory goes beyond this by making a careful distinction between codons that are frequent due to mutational bias and those that are frequent due to translational selection. We have shown that, in a majority of cases, coevolution of tRNAs and codon usage leads to states in which the direction of translational selection is the same as that of the mutational bias. However, there is a substantial minority of cases where selection prefers a different codon from the one that is most frequent under mutation. Our theory explains why these situations are sometimes stable. Therefore we

should not automatically assume that the highest frequency codons are the most preferred by translational selection.

Table 4.1 - Codon usage and tRNA content in A+G families

tRNA gene copies		Number of Cases	Number of cases where each codon is preferred		Low exp genes		high exp genes	
N_U	N_C	Total	#A	#G	π_A	π_G	φ_A	φ_G
1	0	61	45	16	0.736	0.264	0.787	0.213
2	0	21	20	1	0.715	0.285	0.833	0.167
3	0	16	12	4	0.75	0.25	0.814	0.186
4	0	15	14	1	0.684	0.316	0.788	0.212
5	0	8	6	2	0.675	0.325	0.748	0.252
6	0	5	4	1	0.685	0.315	0.729	0.271
7	0	2	1	1	0.703	0.3	0.686	0.314
4	1	1	1	0	0.943	0.057	1	0
7	2	1	1	0	0.696	0.304	0.735	0.265
3	1	12	9	3	0.695	0.305	0.746	0.254
2	1	31	20	11	0.642	0.358	0.664	0.336
4	2	1	1	0	0.673	0.327	0.721	0.279
3	2	1	1	0	0.592	0.408	1	0
1	1	136	52	84	0.459	0.541	0.448	0.552
2	2	4	0	4	0.489	0.511	0.372	0.628
1	2	6	1	5	0.248	0.752	0.083	0.917
1	3	6	0	6	0.209	0.791	0.099	0.901
0	1	2	0	2	0.46	0.54	0.36	0.64
0	2	1	0	1	0.049	0.951	0	1

Table 4.2 – Codon usage in four-codon families where only wobble-U tRNAs are present

tRNA gene copies	Number of cases	Number of cases where each codon is preferred				Low exp genes				High exp genes			
N_U	Total	#U	#C	#A	#G	π_U	π_C	π_A	π_G	φ_U	φ_C	φ_A	φ_G
1	33	13	1	14	5	0.461	0.065	0.408	0.066	0.473	0.039	0.437	0.052
2	18	5	0	12	1	0.385	0.064	0.445	0.100	0.387	0.005	0.563	0.045
3	16	8	0	7	1	0.370	0.113	0.345	0.173	0.487	0.033	0.417	0.063
4	6	5	0	1	0	0.406	0.107	0.348	0.139	0.570	0.030	0.342	0.058
5	2	2	0	0	0	0.403	0.100	0.341	0.156	0.618	0.024	0.295	0.062
6	3	2	0	1	0	0.474	0.135	0.320	0.071	0.572	0.036	0.353	0.040
7	1	0	0	1	0	0.466	0.112	0.338	0.083	0.521	0.034	0.389	0.056

Table 4.3 – Codon usage in four-codon families with both wobble-U and wobble-G tRNAs. Cases are listed in order of increasing ratio of N_G:N_U. In the upper half, $N_G < N_U$. In the lower half $N_G \geq N_U$.

tRNA gene copies		Number of cases	Number of cases where each codon is preferred				Low exp genes				High exp genes			
N_G	N_U	Total	#U	#C	#A	#G	π_U	π_C	π_A	π_G	φ_U	φ_C	φ_A	φ_G
1	5	5	5	0	0	0	0.270	0.200	0.302	0.228	0.494	0.042	0.334	0.130
2	10	1	0	0	1	0	0.295	0.157	0.251	0.298	0.278	0.005	0.624	0.093
1	4	15	14	0	1	0	0.276	0.169	0.300	0.255	0.523	0.051	0.321	0.105
2	7	1	0	0	1	0	0.254	0.224	0.172	0.351	0.302	0.005	0.545	0.148
1	3	20	10	0	10	0	0.344	0.129	0.376	0.152	0.443	0.031	0.466	0.061
2	5	4	4	0	0	0	0.221	0.272	0.175	0.332	0.555	0.107	0.232	0.105
1	2	19	13	0	5	1	0.364	0.151	0.321	0.163	0.494	0.035	0.388	0.083
2	4	5	5	0	0	0	0.252	0.279	0.168	0.302	0.624	0.064	0.239	0.073
4	7	1	1	0	0	0	0.281	0.052	0.586	0.081	0.403	0.052	0.541	0.004
2	3	10	10	0	0	0	0.284	0.249	0.207	0.259	0.540	0.122	0.223	0.114
1	1	64	37	9	9	9	0.405	0.175	0.269	0.151	0.475	0.149	0.262	0.115
2	2	9	9	0	0	0	0.289	0.224	0.271	0.217	0.522	0.150	0.258	0.070
3	3	2	2	0	0	0	0.231	0.314	0.261	0.194	0.463	0.248	0.220	0.068
4	4	1	1	0	0	0	0.371	0.125	0.363	0.141	0.638	0.155	0.194	0.012
4	3	2	2	0	0	0	0.215	0.268	0.313	0.203	0.500	0.165	0.284	0.051
3	2	1	1	0	0	0	0.358	0.217	0.291	0.135	0.553	0.176	0.241	0.030
2	1	3	3	0	0	0	0.316	0.280	0.241	0.163	0.558	0.206	0.192	0.044
3	1	1	1	0	0	0	0.433	0.294	0.164	0.108	0.784	0.170	0.038	0.007
6	2	1	1	0	0	0	0.388	0.363	0.118	0.131	0.659	0.327	0.014	0.000
7	2	1	1	0	0	0	0.391	0.380	0.103	0.127	0.709	0.280	0.010	0.000
4	1	1	1	0	0	0	0.436	0.247	0.151	0.166	0.800	0.142	0.039	0.018
11	2	1	1	0	0	0	0.447	0.358	0.102	0.092	0.683	0.309	0.009	0.000
6	1	1	1	0	0	0	0.380	0.389	0.103	0.128	0.654	0.316	0.020	0.011

Table 4.4 – Codon uses in four codon families with combinations of wobble-C tRNAs with wobble-U and wobble-G tRNAs.

tRNA gene copies			Number of cases	Number of cases where each codon is preferred				Low exp genes				High exp genes			
N_G	N_U	N_C	Total	#U	#C	#A	#G	π_U	π_C	π_A	π_G	φ_U	φ_C	φ_A	φ_G
0	3	1	1	0	0	1	0	0.449	0.046	0.440	0.064	0.363	0.004	0.617	0.016
0	2	1	2	2	0	0	0	0.218	0.203	0.323	0.256	0.467	0.135	0.359	0.041
1	4	1	1	1	0	0	0	0.359	0.221	0.280	0.141	0.398	0.184	0.276	0.142
1	3	1	4	3	0	1	0	0.218	0.291	0.254	0.237	0.370	0.223	0.297	0.110
1	2	1	12	10	0	2	0	0.288	0.201	0.261	0.251	0.442	0.124	0.297	0.136
2	3	1	3	3	0	0	0	0.166	0.387	0.118	0.330	0.336	0.380	0.099	0.186
1	1	1	137	60	41	7	29	0.188	0.334	0.146	0.332	0.208	0.259	0.117	0.317
2	2	1	1	1	0	0	0	0.053	0.440	0.126	0.382	0.164	0.447	0.131	0.257
2	1	1	24	21	2	0	1	0.140	0.471	0.083	0.307	0.268	0.520	0.034	0.177
4	2	1	1	1	0	0	0	0.308	0.194	0.398	0.100	0.652	0.113	0.215	0.020
3	1	1	9	7	2	0	0	0.154	0.525	0.078	0.243	0.265	0.554	0.028	0.152
4	1	1	3	3	0	0	0	0.258	0.502	0.086	0.153	0.576	0.402	0.009	0.014
1	2	2	1	0	0	0	1	0.249	0.236	0.171	0.343	0.181	0.043	0.162	0.614
1	1	2	8	3	1	0	4	0.126	0.225	0.079	0.570	0.141	0.219	0.048	0.592
2	1	2	2	2	0	0	0	0.129	0.378	0.104	0.390	0.334	0.449	0.027	0.190
3	1	2	1	1	0	0	0	0.016	0.545	0.031	0.409	0.059	0.650	0.003	0.287
1	5	3	1	0	0	1	0	0.196	0.228	0.135	0.441	0.180	0.024	0.302	0.494
3	1	3	1	1	0	0	0	0.022	0.527	0.044	0.407	0.088	0.650	0.011	0.252
1	1	4	2	0	0	0	2	0.144	0.135	0.056	0.665	0.055	0.040	0.002	0.902

Table 4.5 Codon usage in four codon families involving wobble-A (or I) tRNAs.

tRNA gene copies				Number of cases	Number of cases where each codon is preferred				Low exp genes				High exp genes			
N_A	N_G	N_U	N_C	Total	#U	#C	#A	#G	π_U	π_C	π_A	π_G	φ_U	φ_C	φ_A	φ_G
1	0	0	0	1	1	0	0	0	0.122	0.548	0.068	0.262	0.262	0.685	0.008	0.044
1	0	0	1	17	14	1	0	2	0.223	0.427	0.113	0.237	0.368	0.473	0.045	0.115
2	0	0	2	1	1	0	0	0	0.263	0.520	0.074	0.142	0.634	0.247	0.004	0.016
2	0	0	1	16	16	0	0	0	0.427	0.321	0.140	0.112	0.687	0.289	0.011	0.013
3	0	0	1	4	4	0	0	0	0.509	0.197	0.199	0.096	0.813	0.172	0.014	0.001
4	0	0	1	5	5	0	0	0	0.310	0.400	0.124	0.167	0.672	0.314	0.006	0.007
6	0	0	1	3	3	0	0	0	0.424	0.384	0.126	0.065	0.762	0.231	0.005	0.002
8	0	0	1	1	1	0	0	0	0.469	0.343	0.163	0.026	0.779	0.215	0.006	0.000
1	0	1	0	10	8	0	1	1	0.570	0.147	0.213	0.071	0.737	0.089	0.107	0.066
1	0	2	0	2	2	0	0	0	0.418	0.197	0.230	0.155	0.922	0.033	0.038	0.007
1	0	3	0	1	1	0	0	0	0.476	0.138	0.285	0.101	0.935	0.027	0.031	0.007
1	0	1	1	2	2	0	0	0	0.158	0.407	0.170	0.264	0.480	0.415	0.051	0.053
2	0	2	1	1	1	0	0	0	0.561	0.163	0.163	0.114	0.792	0.205	0.000	0.003
1	1	1	1	1	1	0	0	0	0.246	0.223	0.278	0.353	0.537	0.151	0.141	0.171

Table 4.6 Summary of the modifications of different bases those are found so far. There are not many experiments on base modifications. However, this table already shows that the types of modification are very diversified but can be grouped into just one or two category for each base (also refer to Fig 4. 2 and text for details). The function of modification is to enhance the translation speed or accuracy or both.

BASE	G	U	U	A
TYPE OF MODIFICATION	Queuosine	xm^5U	xmo^5U	I
CODON FAMILLIES	U+C families for Tyr, His, Asn, Asp	A+G families	Most four codon families	Four codon family CGN for Arg
DOMAINS IN WHICH MODIFICATION IS FOUND	all domains	all domains	bacteria	bacteria and eucaryote
FUNCTION	Enhance the pairing of Q-U	Prevent the parings of U-U and U-C	Expand/enhance pairing	Allowing I pairs with U,C,A
EFFECTS	Speed	accuracy	Speed	?Speed
PREFERRED CODON	C	A	U, A	U

Chapter 5

Translational accuracy

Preface

The effect of translational accuracy on codon usage bias is studied. Two hypotheses of translational accuracy are tested and the results show that the effect is rather weak across all bacterial genomes.

5.1 Introduction

In chapters 3 and 4, a model for the effect of translation speed on codon usage bias has been studied. However, codon bias might also arise due to selection for translational accuracy, *i.e.*, the preferred codons would be those which are most accurately translated, not those which are fastest. Generally the effect of translational accuracy can be tested by the following two hypotheses. At some sites in proteins, amino acids are highly conserved across many species. It is presumed that these sites are conserved because they are particularly important for protein functions. Thus, accurate translation of these sites should be particularly important. This is called missense error. If this hypothesis is true, preferred codons should be more frequent in these conserved sites. Akashi did this test first in *Drosophila* and selection due to accuracy was found (Akashi, 2003). Another hypothesis is called nonsense error hypothesis (Eyre-Walker *et al.*, 2006). During translation, ribosomes will move along the mRNA chain and an amino acid chain will be

produced. However, sometimes this process will be terminated earlier. This premature termination is a waste of energy. So it is an advantage to avoid a high rate of premature termination. According to this hypothesis, preferred codons are those that are least likely to terminate prematurely. This leads to two expectations. First, the longer the gene sequence, the more accurate the sequence should be. Hence there should be a higher frequency of preferred codons in longer genes. Second, within a gene sequence the accuracy on the 3' half should be higher than that on the 5' half. If this hypothesis is true, the codon bias may be higher in the 3' half of each sequence, and higher in the longer sequences. Eyre-Walker *et al.* have discussed all these hypotheses in three *Escherichia coli* genomes. In that paper they also tested missense error. In this thesis, only the hypotheses of conserved sites and premature terminations are tested. A weak but significant effect of the translation accuracy on codon usage bias is found in most bacteria. Another interesting result is that the C codon is selected on the grounds of both speed and accuracy in the U+C codon families. We have not yet examined the other cases, but it could be a general rule that the fast codon is also more accurate.

The effect of accuracy has been tested just in a few species before this work. It is the first time that a systematic analysis is preformed on both effects of translation accuracy and speed across all bacteria.

5.2 Sequence data

In Chapter 3 and 4, the 80 species used in Sharp's paper (2005) are used to test speed hypothesis. Now we will look at the codon usage in high expression genes only to test the

hypothesis of translation accuracy. For this analysis, the sequence data were obtained as follows: Firstly, 8 extra species were randomly chosen and added which represent species from groups that were not included in the original 80 species used by Sharp: *Bordetella bronchiseptica RB50, Bordetella pertussis Tohama I, Anaeromyxobacter dehalogenans, Desulfovibrio desulfuricans, Geobacter sp. FRC-32, Syntrophobacter fumaroxidans MPOB, Pelobacter carbinolicus DSM 2380, Magnetococcus sp. MC-1* (see Appendix). 54 ribosomal protein genes and 3 elongation factors of *Escherichia coli* were downloaded from the *Escherichia coli* database. The BLAST was performed with *Escherichia coli* K-12 protein sequence as query genes to find out the corresponding genes in the other species. Both the protein sequences and DNA sequences were downloaded. We did BLAST with protein sequence because it was easier to find matching proteins in divergent species than to find matching DNA sequences. Then the protein sequences of all genes of all species were aligned with MUSCLE. Secondly, the DNA sequences were aligned with the protein sequences as a reference. Only 47 genes out of all 57 genes were selected for further analysis according to two rules. 1) Each sequence of any ribosomal protein gene or elongation factor should have a similar length using each corresponding gene of *Escherichia coli* K-12 as a reference. A gene was excluded if it was more than two times longer. This is reasonable because the much longer sequences maybe due to some special function of ribosomes of some species, which is quite different from that of the rest species. Then there is no point to compare this much longer sequence with the rest sequences. Another reason may be annotation errors on the websites. 2) A gene were excluded if the E values of BLAST was larger than 0.05. As we know, only small E

values mean the match of alignment is highly significant. So with these rules, we have made a reliable alignment of both proteins and DNAs. The codon usage can then be calculated by just counting the codon numbers on the conserved and variable sites, which will be defined in next section in detail. We list all 47 high expression genes here: L1-L7/L12, L9-L11, L13-L22, L24, L27, L28, L31, L35, S2-S20, EF-G, EF-Tu, EF-Ts.

5.3 Definition of conserved and variable sites

Examples of real sequence alignments are shown in Fig. 5.1 and Fig. 5.2. The conserved and variable sites of the protein sequences are defined as follows. In Fig. 5.1 there is a Lys (K) at position 5 in all species in this alignment. However there are rather few sites where an amino acid is conserved in every species. Therefore we need a less strict definition of what counts as conserved. We define a threshold percentage ε. If the most frequent amino acid at a site has a frequency greater than or equal to ε, this site is counted as conserved. If the most frequent amino acid has frequency less than ε, the site is counted as variable. Codons for the most frequent amino acid at conserved sites are counted as conserved. Codons for minority amino acids at the conserved site are counted as variable. Codons for all amino acids at variable sites are counted as variable. Having defined which codon positions are conserved and variable, we sum up the number of occurrences of each codon at conserved and variable position in that species. We then perform the following tests to see whether codon usage is different between conserved and variable sites. Here the alignment is used to determine which sites are conserved, but the counting of conserved and variable sites is done on each species separately.

```
                        *         20         *
ESCCOL      : MAVVKCKPTSPGRRHVVKVVNPELHKGKPFAPLLEKNS :  38
SALTYP      : MAVVKCKPTSPGRRHVVKVVNPELHKGKPFAPLVEKNS :  38
YERPES      : MAIVKCKPTSPGRRHVVKVVNPELHKGKPYAPLLEKLS :  38
BUCAPH      : MAVVKCKPTSPGRRHVIKVVNNELYKGKPYSLLLRKKS :  38
WIGGLO      : MTLNKCNPTTPSRRHTVKVVNNKLYKGKSFFKLTKSLN :  38
HAEINF      : MAIVKCKPTSAGRRHVVKIVNPELHKGKPYAPLLDTKS :  38
PASMUL      : MAIVKCKPTSAGRRHVVKIVNPELHKGKPYAPLLDTKS :  38
VIBCHO      : MAIVKCKPTSAGRRHVVKVVNADLHKGKPYAPLLEKNS :  38
VIBPAR      : MAIVKCKPTSPGRRHVVKVVNADLHKGKPYAPLLEKNS :  38
VIBVUL      : MAIVKCKPTSAGRRHVVKVVNADLHKGKPYAPLLEKNS :  38
SHEONE      : MAVIKCKPTSPGRRHVVKVVNSDLHKGKPFAGLLAKKS :  38
PSEAER      : MAIVKCKPTSAGRRFVVKVVNQELHKGAPYAPLLEKKS :  38
PSEPUT      : MAIVKCKPTSPGRRFVVKVVNKELHKGAPHAPLLEKKS :  38
PSESYR      : MAIVKCKPTSPGRRFVVKVVNQELHKGAPHAPLLEKKS :  38
XANAXO      : MPLMKFKPTSPGRRSAVRVVTPDLHKGAPHAALLDSQS :  38
XANCAM      : MPLMKFKPTSPGRRSAVRVVTPDLHKGAPHAALLDSQS :  38
XYLFAS      : VPLIKFKPTSPGRRSAARVVTPNIHKGSPHASLLESQS :  38
COXBUR      : MALVKTKPTSPGRRFVVKVVHPELHKGDPYAPLVESKN :  38
              6a6 KckPT3 gRR v 46Vn  6hKG p a L    s
```

Fig. 5.1 Alignment of protein sequences of ribosomal protein L2.

```
                        *         20         *
ESCCOL      : ATGGCAGTTGTTAAATGTAAACCGACATCTCCGGGTCG :  38
SALTYP      : ATGGCAGTTGTTAAATGTAAACCGACATCTCCGGGTCG :  38
YERPES      : ATGGCAATTGTTAAATGTAAACCTACGTCTCCGGGTCG :  38
BUCAPH      : ATGGCAGTTGTTAAGTGTAAACCGACATCTCCGGGTCG :  38
WIGGLO      : ATGACATTAAATAAATGTAACCCAACAACTCCTAGTCG :  38
HAEINF      : ATGGCTATCGTTAAATGTAAGCCGACCTCCGCTGGTCG :  38
PASMUL      : ATGGCTATCGTTAAATGTAAGCCGACCTCCGCTGGTCG :  38
VIBCHO      : ATGGCTATTGTTAAATGTAAGCCGACTTCGGCTGGTCG :  38
VIBPAR      : ATGGCTATTGTTAAATGTAAGCCGACTTCCCCTGGTCG :  38
VIBVUL      : ATGGCTATTGTTAAATGTAAGCCGACTTCGGCTGGTCG :  38
SHEONE      : ATGGCAGTTATTAAGTGTAAGCCAACCTCTCCAGGTCG :  38
PSEAER      : ATGGCAATCGTTAAGTGCAAACCGACTTCCGCTGGTCG :  38
PSEPUT      : ATGGCAATCGTTAAATGCAAACCGACTTCCCCTGGCCG :  38
PSESYR      : ATGGCAATCGTTAAATGCAAACCGACTTCCCCTGGCCG :  38
XANAXO      : ATGCCATTGATGAAGTTCAAACCCACCTCTCCCGGCCG :  38
XANCAM      : ATGCCATTGATGAAGTTCAAACCCACCTCTCCCGGCCG :  38
XYLFAS      : GTGCCGTTAATAAAATTCAAACCCACTTCTCCAGGTCG :  38
COXBUR      : ATGGCTTTAGTAAAAACAAAACCAACGTCCCCAGGGCG :  38
              aTGgC  T  ttAA tg AA CC AC tC  C gG CG
```

Fig 5. 2 Alignment of DNA sequences of ribosomal protein L2. This alignment is aligned using the protein sequence in Fig. 5.1 as a reference.

5.4 Simple χ^2 test for codon bias in conserved versus variable sites

Since there are relatively small numbers of codons for each amino acid in the sets of high-expression genes, statistical noise could be large. The codon counts for the conserved and variable codons form a 2 × 2 table. We used a χ^2 test to test the null hypothesis that the codon frequencies are the same in the two groups of genes in each set. In above section we define a threshold percentage ε. However, there is no general rule that which threshold percentage is the best. So we used different threshold percentage ε=60%, 70%, 80% and 90%. In U+C codon families the codon data with these threshold percentages for conserved sites were analyzed respectively and it was found that only 13.6%, 12.6%, 15.6% or 12.1% of codons showed a significant codon bias between the different categories of conserved and variable sites in high expression genes (Table 5.1). In comparison, if the same χ^2 test is performed to check for difference between the high and low expression genes, 80.4% of U+C codon families show significant difference.

A statistical analysis is also performed by counting how many cases where the selection directions due to accuracy or speed are the same or not in U+C codon families. In the cases where both effects of accuracy and speed are significant, the selection pressure goes in the same direction in most cases. Table 5.2 shows that C codon is preferred in 55 of 65 cases where both tests are significant. This suggests that C codon is preferred both on speed and accuracy. This is quite interesting since it means the fast codon is also the accurate one. Scientists have kept arguing that CUB is due to either speed or accuracy. However, a natural conclusion could be that CUB is due to speed and accuracy at the same time. CUB is the consequence of optimization of both translation

speed and accuracy, which means organisms take both advantages during the evolution to make them fit the environment better. Our conclusion is, therefore, that selection on accuracy is relatively a weak effect, and weak selection for accuracy appears to be working in combination with selection for speed.

Table 5.1 χ^2 test on all three sets of genes. Statistical significant cases of all amino acids in all species for different percentage of conservations. The first row is for the sets of high and low expression genes. The last row is for the gene sets of 3' and 5' halves.

	SIG C	SIG U	NS
HL	367 (71.9%)	48(9.3%)	101(19.6%)
CV 90%	54(10.5%)	8(1.6%)	454(88.0%)
CV 80%	69(13.4%)	6(1.2%)	441(85.5%)
CV 70%	58(11.2%)	7(1.4%)	451(87.4%)
CV 60%	62(12%)	8(1.6%)	446(86.4%)
Two Halves	36(7%)	9(1.7%)	471(91.3%)

Table 5.2 Comparison of details of χ^2 test for conserved and variable sites with a percentage of 80 and for high and low expression genes in all species.

		Conserved and Variable with 80%			
		C	U	NS	sum
High and Low	C	55	5	307	367
	U	5	0	43	48
	NS	9	1	91	101
	sum	69	6	441	

5.5 Simple χ^2 test of nonsense error

The hypothesis of nonsense error is also tested with 47 high expression genes in the same way where each gene sequence is divided to two halves. Codon frequencies are counted in these two halves. A χ^2 test is then preformed in U+C codon families in all species and 8.7% of codon families showed a significant codon bias between 3' and 5' halves of high expression genes, which indicates a fairly weak selection against premature termination.

So all tests on conserved versus variable sites in previous section and 3' half versus 5' half in this section show a larger percentage (p>5%), which clearly says that the effect is significant and cannot be statistic noise. However, this is a weak effect compared with the test of speed hypothesis (significance with a percentage of 80.4). Another possibility is that the data in accuracy test are less since we divide the high expression genes to two sets. So we would like to have information from all 59 codons to be considered, instead of 12 U+C codons. This is done by a new method in next section.

5.6 A method to test the effect of selection in all codon families

Here we develop a more powerful method which is used to test the effect of selection in all amino acids. All codons are separated to two groups A and B, which can be high expression genes and low expression genes, or conserved sites and variable sites, or the 3' half of a gene and 5'half of that gene respectively. Let ϕ_i^A and ϕ_i^B be the numbers of occurrences of codon i in groups A and B. Then the relative frequencies of codons in each group and in the combined data from both groups are

$$\phi_i^A = \frac{n_i^A}{\sum_{aa} n_i^A}$$
$$\phi_i^B = \frac{n_i^B}{\sum_{aa} n_i^B}$$
$$\phi_i^0 = \frac{n_i^A + n_i^B}{\sum_{aa} (n_i^A + n_i^B)} \qquad (5.1)$$

Here $\sum_{aa}$ means the sum is over codons for one particular amino acid.

We can now use the maximum likelihood (ML) method to develop a statistical test for difference in frequencies between groups. In the null model, we suppose the frequencies in both groups are same. The ML estimators of the frequencies are equal to the observed frequencies ϕ_i^0. The log likelihood according to the null model is

$$\ln L_0 = \sum_i \left(n_i^A + n_i^B\right) \ln \phi_i^0 \qquad (5.2)$$

We then consider an alternative model in which codon frequencies are allowed to be different in the two groups. The ML estimators of the frequencies are then given by ϕ_i^A and ϕ_i^B and the log likelihood is

$$\ln L_1 = \sum_i \left(n_i^A \ln \phi_i^A + n_i^B \ln \phi_i^B\right) \qquad (5.3)$$

The null model is a nested model within the alternative model. Therefore we can use the likelihood ratio test (Whelan and Goldman, 2001) to determine whether the alternative model is a significant improvement on the null model. If the null model is true, the quantity 2Δ defined below should have a χ^2 distribution.

$$2\Delta = 2\ln(\frac{L_1}{L_0})$$
$$= 2\left(\sum_i n_i^B \ln(\frac{\phi_i^B}{\phi_i^0}) + \sum_i n_i^A (\frac{\phi_i^A}{\phi_i^0})\right) \quad (5.4)$$

The values of 2Δ can be used to test for deviations from the null, *i.e.* to show that there is significant difference between two groups.

It is useful to define quantities δ_A and δ_B that measure the improvement in the log likelihood per codon. In cases where the null is not true, these parameters measure the strength of the deviation in codon usage between the two sets.

$$\delta_A = \sum_i n_i^A \ln \frac{\phi_i^A}{\phi_i^0} \Big/ \sum_i n_i^A ,$$
$$\delta_B = \sum_i n_i^B \ln \frac{\phi_i^B}{\phi_i^0} \Big/ \sum_i n_i^B \quad (5.5)$$

The δ is a powerful parameter to represent the selection acting on codon usage. It has many advantages compared to CAI, N_C and S because: 1) It is a measurement of any two groups where selection or difference exists. 2) CAI just considers the effect from high expression genes, while δ, from both high and low. 3) N_C cannot distinguish the bias caused by mutation from selection. 4) The S values were calculated just in U+C codon families in Sharp's paper (2005) and expanded to all codons in our papers (Ran and Higgs, 2010; Higgs and Ran, 2008). However, it is hard to do a statistical test on all codons together, while for δ values, the maximum likelihood ratio 2Δ is a good test and easy to perform.

5.7 Comparison of strength of bias δ for different hypotheses

The strength of bias δ was calculated first in the codons of high and low expression genes. The result is shown in Fig. 5.3. A strong and significant correlation exists between strength of bias δ_H in high expression genes and growth rate (1/minimum doubling time). The linear fitting of δ_H is done and the slope is 0.092, R is 0.71 and $p<0.0001$. However, No correlation is found between δ_L and growth rate. This means that there is strong selection existing in high expression genes and can be interpreted to be due to translation speed. δ_L is small in all cases because the number of codons in the low expression genes is much larger than in the high expression genes, *i.e.* $\phi_i^L \approx \phi_i^0$. Then the same performance was done on the codons of conserved and variable sites, and the codons of 3' half and 5' half. However, no correlation is found between δ_C, δ_V, and growth rate in Fig. 5.4, or between $\delta_{3'half}$, $\delta_{5'half}$ and growth rate in Fig. 5.5. This means the four strengths of bias have nothing to do with translation speed and have to be related to translation accuracy only. The fact that the values of these four parameters are at least an order smaller than that of δ_H makes it clear that the selection due to accuracy is very weak, whether from the view of nonsense or missense error test..

The statistic tests of 2Δ for different groups of codons were performed and the percentages of significance are counted. For codons in high and low expression genes, the test shows all 2Δ values are all significant. For codons in conserved sites and variable sites with a threshold ε=80%, and 3' half and the 5' half, δ values are all significant with a percentage of 70.9 and 46.5 respectively. The significance is much higher than that in the statistic test in 6 U+C families (12 codons) because much more data (59 codons) are taken into account. So the maximum likelihood ratio test 2Δ is a more powerful statistic

test. Although the effect is statistic significant with this new method for conserved versus variable sites and 3' half versus 5' half, the selection due to accuracy is rather weak if we compare their δ values with those from high and low expression gene sets. The later one is at least an order of magnitude larger. So the conclusion is that the tests on the two accuracy hypotheses show the selection for accuracy is a weak effect acting in combination with the selection for speed. This conclusion is the same as that in section 5.4.

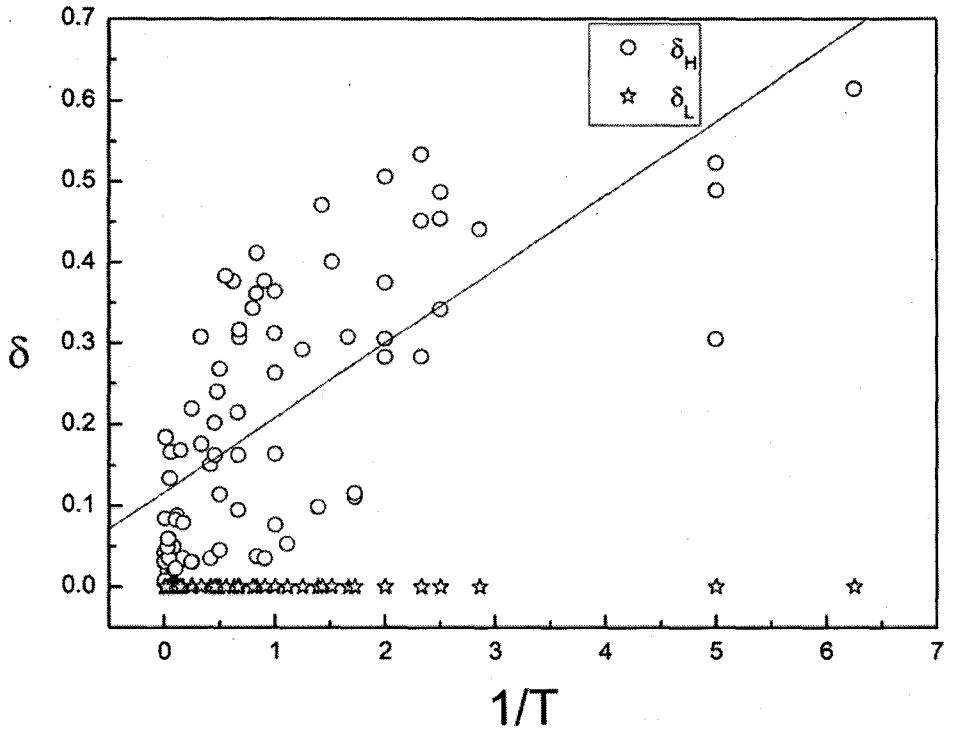


Fig. 5.3 A strong and significant correlation exists between selection parameter δ_H in high expression genes and growth rate (1/minimum doubling time). The line is the fitting of δ_H. The slope is 0.092, R is 0.71 and $p<0.0001$. No correlation is found between δ_L and growth rate.

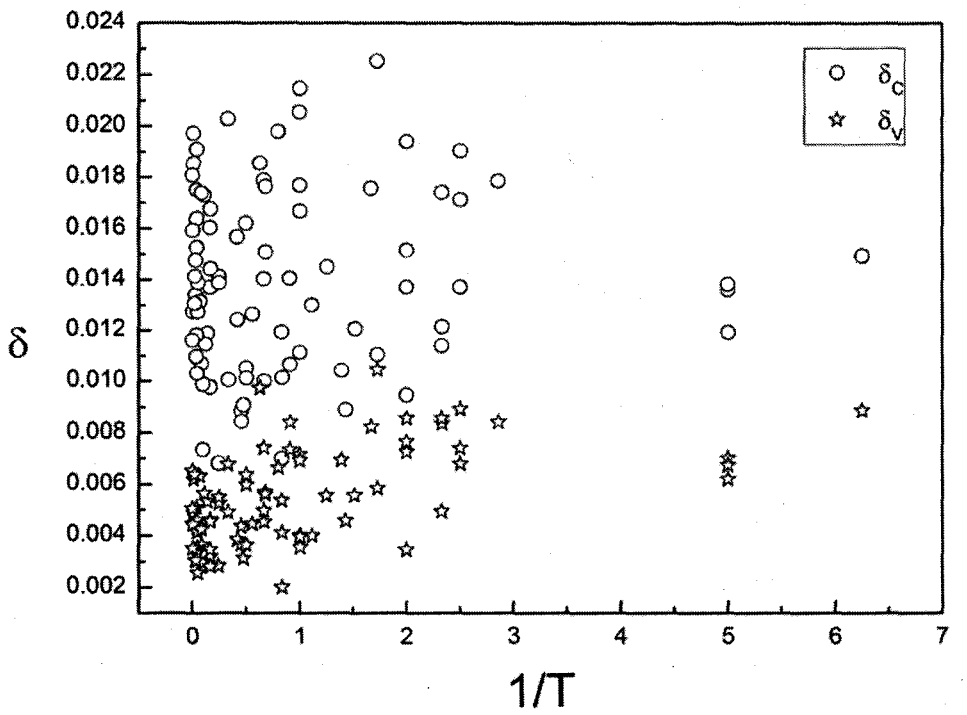


Fig. 5.4 No correlation is found for δ_C, δ_V and growth rate. The values of δ_C and δ_V and the difference between δ_C and δ_V are small, which means only a weak selection exists in the conserved sites.

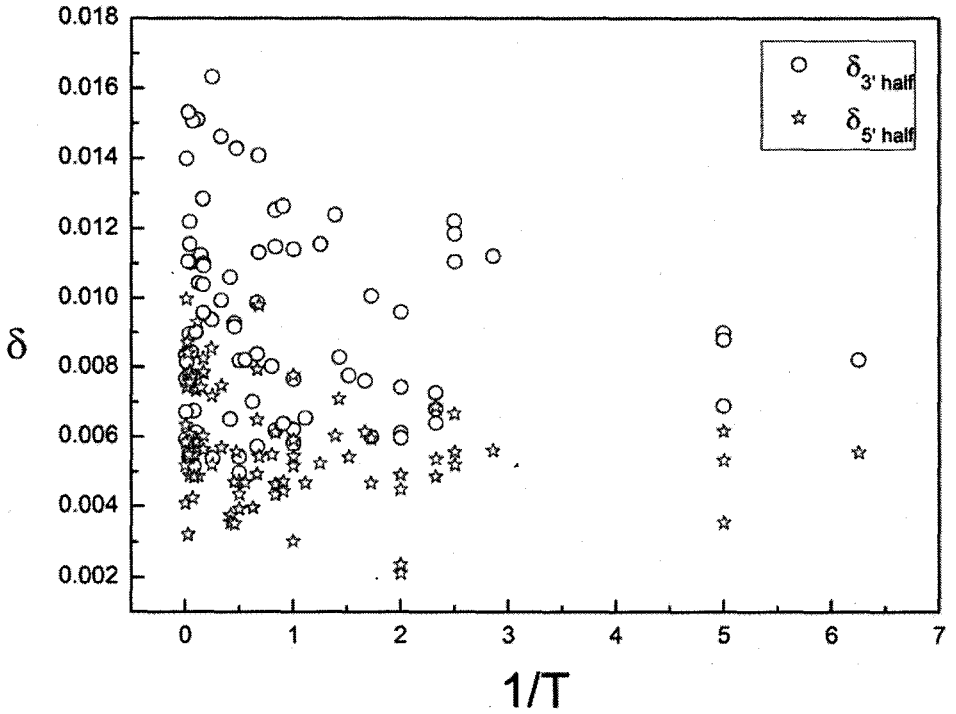


Fig. 5.5 No correlation is found between growth rate and the δ values of the 3'half or 5' half of genes in bacteria. The values of $\delta_{3',\,half}$ and $\delta_{5',\,half}$ and the difference between $\delta_{3',\,half}$ and $\delta_{5',\,half}$ are small, which means only a weak selection exists in the two halves. In other word, nonsense error is weak.

Chapter 6

Discussion, Conclusion and Future work

6.1 Discussion of relative importance of speed and accuracy

In chapter 3 and 4, the model of translational selection that we have used to interpret the codon usage data is based on the assumption that translational speed is the key factor. In chapter 5 the effect of translational accuracy on CUB is studied – *i.e.* the preferred codons are those for which the error rate is smallest rather than those which are most rapidly translated. Now we will discuss both possible causes of selection. Firstly we will argue that there are important aspects of these data that can be most easily explained in terms of selection for speed, although it is quite possible that selection for accuracy is operating at the same time.

Experimental evidences show that there is significant difference in translation speeds between synonymous codons. Curran and Yarus (1989) and Sorensen and Pedersen (1991) found that codons that were preferred in *E. coli* according to codon usage data were indeed translated faster. It is also known that insertion of blocks of slow codons into a sequence has a significant effect on protein production rate (Mitarai *et al.* 2008) and that these effects can be well described by a model that considers the position of the fast and slow codons along the mRNA. In section 1.7 I introduced the experiments which have attempted to measure the rates of the different steps involved in the translation cycle for each codon. However, it is not yet clear exactly which of the underlying steps leads to variation in the effective rate. Also, different groups use

different kinetic schemes, as pointed out by Ninio (2006), so there is not yet complete agreement on what the underlying steps are. We wish to emphasize that selection on codon bias shows up in the simplest possible case where there is only one tRNA that pairs with two synonymous codons. Thus, if it is speed that is under selection, there must be a difference in the rate at which the ribosome recognizes and processes the tRNA that depends on the details of the codon-anticodon interaction, which has been found by Blanchard (2004), though he didn't work on the difference among synonymous codons.

Accuracy-based arguments assume that mistranslation has a cost because some mistranslated proteins are non-functional (Drummond and Wilke, 2008). Selection for accuracy can explain observed differences in codon usage between conserved and variable sites (Akashi, 1994; Stoletzki and Eyre-Walker, 2006) seen in some species, which we would not expect from selection for speed alone. In general, the probability of mistranslating a codon is $m_i/(m_i + r_i)$, where r_i is the rate of correct translation and m_i is the rate of incorrect translation. If accuracy is the key factor, then in order to understand this fully, it will be necessary to measure the mispairing rates of each tRNA with all the non-cognate codons as well as the correct pairing rates with the cognate codons, which has never been done yet. The relative accuracy of synonymous codons has been measured in a few cases (Precup and Parker, 1987; Kramer and Farabaugh, 2007) although there is no systematic study of relative accuracy that covers all codons in a given organism. An interesting special case related to selection against inaccurate codons is the elimination of ambiguously translated codons during periods of codon reassignment, as has been observed with *Candida* (Butler *et al.* 2009).

If a tRNA is rare, then the rate of translation of its cognate codons will be slow, but it is also likely to be inaccurate because the ratio m_i/r_i will tend to be larger when r_i is smaller. For example, Kramer and Farabaugh (2007) showed that there is a relatively high rate of mistranslation of the AGR Arg codons in *E. coli*, for which the cognate tRNA is rare. This might explain selection between the CGN Arg codons and the AGR codons, but it is less clear that this argument can be used to explain selection between codons translated by the same tRNA, such as any of the U+C or A+G pairs considered in this thesis.

Despite these caveats regarding the possible relevance of accuracy as well as speed, there are two aspects of the data that point to the fundamental importance of translational speed. Firstly, it has been show by Rocha (2004) and in Chapter 3 that codon bias is higher in bacteria with faster growth rates. This is a natural expectation if speed is important – bacteria living in a niche where rapid cell division is advantageous need to adapt to optimize their rate of protein synthesis. It is difficult to see why this correlation should occur if selection were solely due to accuracy. Secondly, it is found that there are more duplicate tRNA genes in bacteria that are rapidly multiplying and in species where the codon bias is strong (Rocha, 2004; Chapter 3, Fig. 3.6). Duplication of a tRNA leads to higher tRNA concentration and hence increases translational speed. In Chapter 3 our theory predicts in which circumstances duplications are favoured by selection for translational speed. In contrast it is not clear that selection for accuracy alone will favour gene duplications. Duplicating one tRNA should increase the rate of translation of its cognate codons and hence also increase their accuracy. However, it will also increase the

rate at which this tRNA mispairs with non-cognate codons. It is not clear which of these is more important. Furthermore, if we make a general duplication of all genes, this will increase all the correct pairing rates and mispairing rates proportionately, so there should be no change in accuracy, but a large increase in speed of all codons.

We emphasize that the above arguments apply only to bacteria, and it may well be that speed is less relevant in multicellular organisms than in bacteria. Several examples where arguments for translational accuracy have been made are multicellular eukaryotes (Drummond and Wilke, 2008). However, eukaryotes tend to have large numbers of tRNA genes, and the number of copies of each type of tRNA is correlated with the codon frequencies, as has been shown, for example, in *Caenorhabditis elegans* (Duret, 2000) and humans (Lavner and Kotlar, 2005). Therefore tRNA copy number and codon frequencies are also coevolving in eukaryotes, and this means that there is still a fundamental role for speed and efficiency even in multicellular organisms

From the results in chapter 5 it is found that the C codon is selected on the grounds of both speed and accuracy in the U+C codon families. We have not yet examined the other cases, but it could be a general rule that the fast codon is also more accurate. This is more reasonable than the disagreements among scientists who tried to stress on just one of them. This makes more sense if we relate this assumption to tRNA modification in Chapter 4 where modification will enhance the translation speed or accuracy or, in many cases, both. The category of modification in section 4.5 is from a few species since only a few experiments have been done to explore the modifications. However, from the results in chapter 5, we can infer that tRNA modifications will

enhance both factors at the same time in most cases in most species from bacteria to drosophila. We make this deduction because the modified Q-tRNA is also found recently in drosophila and its function is to enhance both translation speed and accuracy, which is exactly the same as that in bacteria. Finally, we want to point out that this is the first time that accuracy and efficiency are studied in so many species systemically on their effects on CUB.

6.2 Summary of Conclusions in the thesis

The big question discussed in this thesis is the origin of CUB in bacteria. Our conclusion is that apart from mutation bias, selection for speed is the main cause compared with selection for accuracy.

An organism prefers to use codons that are more rapidly translated because this makes most efficient use of the limited number of ribosomes in the cell. This was supported firstly by experiments showed that preferred codons are also the fast translated codons. In this thesis, selection on CUB is found in most bacteria species by comparison of high and low expression genes. The theory of multiple stable states describes evolution between tRNA copy number and CUB. Fast translation is an advantage for fast multiplying bacteria, which have a large growth rate– therefore they have duplicate tRNAs and stronger codon bias. With our translation model the new parameter K is also used to make predictions on selection direction, which is dependent on the presence of different tRNAs anticodons and tRNA gene copy numbers. So our researches afford solid support for that speed is a cause of CUB.

A further study on codon-anticodon interaction was performed without either speed or accuracy hypothesis. The surprising results is that preferred anticodon-codon pairs are not always Watson-Crick pairs, especially a high interaction of U-tRNA with U-ending codon is found in four codon families. This leads to the finding of origin of tRNA modifications in the light of evolution: For a higher translation speed and/or accuracy, different groups of modifications are introduced for different codon families which follows different translation patterns. These modifications finally influence CUB greatly. With the codon frequencies and selection strengths attracted from sequences, predictions on experimental measurements of anticodon-codon interactions are made, which can be tested in the future.

Mistranslation will lead to an unfinished amino acid chain (nonsense error) or a wrong amino acid (missense error), which is a waste of energy and sometimes even toxic to the cell. So codon usage should be high for the accurate codons to avoid these mistranslations. An effective way is to compare the CUB on conserved and variable sites in the protein and DNA sequences by genome comparison. The conserved ones are supposed to be essential for protein functions and high accuracy of translation should be required. With our bioinformatics work and a new parameter δ, comparison of the effects on CUB from speed and accuracy is performed. However, it is found that the major factor causing selection is speed. The effect of selection for accuracy is much weak and acting in combination with the selection for speed. The method with this new parameter is superior because it makes systemic comparison of selection within and among genomes possible for both speed and accuracy hypotheses. It can also be used to test the difference

or selection in any two groups of gene sets, or position effect in the genomes.

Above research is unique because this is the first time that a systemic analysis of the origin of CUB is preformed in a large number of genomes across all bacteria, taking both effects of speed and accuracy into account. CUB is studied in both qualitative and quantitative ways here.

6.2 Future research

6.2.1 Relative importance of speed and accuracy in different domains

For large multicellular organisms such as mammals, the small doubling time may not be a key factor for their adaption to the environment. Nevertheless, accurate translation may still be important. So the CUB may be affected mainly by accuracy in these organisms. In mouse and human accuracy has been shown to play an important role (Drummond and Wilke, 2008). From our conserved site test and nonsense error test it is already known that the effect of accuracy is fairly weak in bacteria. If this is also true for multicellular organisms, it would be expected that the CUB in most large multicellular organisms could be fairly low. It will then provide a robust test saying that the effect of accuracy on CUB is weak in all domains, while the effect of speed is strong for bacteria and lower eukaryotes. Generally, the effect of speed and accuracy on CUB can be an interesting question in other domains of life.

6.2.2 A kinetic model of Translation process

We discussed the translation process in the introduction where the schemes and experiments are not consistent among different groups of scientists (Rodnina *et al.*, 2001; Blanchard *et al.*, 2004; Ninio, 2006). All these experiments are dealing with the differences of translation rates for cognate (three bases are of a complete match between the codon and anticodon) and near-cognate (Codon position 1st or 3rd doesn't match with anticodon) and non-cognate codons (more than one bases don't match), instead of synonymous codons as we discuss in this thesis. If there are experiments which measure the translation rates of all steps of translation process using synonymous codons, we can use our theories and sequence analysis from all genomes to study these data and a better model of translation kinetics can be built. We notice Zouridis and Hatzimanikatis (2007, 2008) attempted to build such a model. However, their conclusion was that U-ending codon was preferred to C-ending codon in U+C codon families. We believe it is wrong because a wrong defined group of mismatch between codon and anticdon is used in their model.

We strongly suggest a particularly interesting experiment on the measurement of translation rates using the A+G two codon families and four codon families, which should show a big difference for these two groups of codons. Another comparison experiment will also be important, which is to explore the difference of translation rates using a modified tRNA and an unmodified one. tRNA plays such an important role during translation and evolution. Many scientists have done excellent work (For example, Grosjean *et al.*, 2010; Agris, 2004, 2008). However, lots of experiments on tRNA modification should be done in different organisms and domains to push this field ahead.

6.2.3 MCMC simulation on anticodon-codon interactions

The qualitative study on anticodon-codon interaction has been introduced in Chapter 4 where the relative magnitude of these b's are predicted. However, it will be more sensible to have some values that can be tested by future experiments. We have already used a Markov Chain Monte Carlo method to fit these parameters which optimize translation time. The fitted parameters works well for two codon families and are consistent with Watson-Crick pair rules. However, the b_{UU} is large in four codon families mainly because we didn't find a proper constraint on that parameter. So the next step should be making a further investigation of the experiments of translation kinetics and finding some constraint to do the simulation again.

References

Agris PF. 2004. Survey and summary decoding the genome: a modified view. Nucleic Acids Res. 32: 223-238.

Agris PF. 2008. Bringing order to translation: the contributions of transfer RNA anticodon-domain modifications. EMBO Report 9:629-635.

Akashi H. 1994. Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. Genetics 164: 1291-1303.

Akashi H. 2003. Translational selection and yeast proteome evolution. Genetics 164: 1291-1303.

Ames B and Hartmann P. 1963. The histidine operon. Cold Spring Harbor Symposium Quantitative Biology 28: 349-356.

Andachi Y, Yamao F, Muto A, Osawa S. 1989. Codon recognition patterns as deduced from sequences of the complete set of transfer RNAs species in *Mycoplasma capricolum*. J. Mol. Biol. 209: 37-54.

Ardell DH and Kirsebom LA. 2005. The genomic pattern of tDNA operon expression in *E. coli*. PLoS Comp. Biol. 1(1):e12.

Arnold RJ and Reilly JP. 1999. Observation of Escherichia coli ribosomal proteins and their posttranslational modifications by mass spectrometry. Anal Biochem. v269 (1): pp 105-12.

Ashraf SS, Sochacka E, Cain R, Guenther R, Malkiewicz A, Agris PF. 1999. Single atom modification (O→S) of tRNA confers ribosome binding, RNA 5: 188–194.

Benzécri J.-P. 1973. *L'Analyse des Données. Volume II. L'Analyse des Correspondences*. Paris, France: Dunod.

Bernstein JA, Khodursky AB, Lin PH, Lin-Chao S, Cohen SN. 2002. Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays. PNAS 99: 9697-9702.

Blanchard SC, Gonzalez RL Jr., Kim HD, Chu S, Puglisi JD. 2004. tRNA selection and kinetic proofreading in translation. Nature Struct. Mol. Biol. 11:1008-1014.

Boren T, Elias P, Samulelsson T, Claesson C, Barciszewska M, Gehrke CW, Kuo KC, Lustig F. 1993. Undescrimitating codon reading with adenosine in the wobble position. J. Mol. Biol. 230: 739-749.

Bulmer M. 1987. Coevolution of codon usage and transfer RNA abundance. Nature 325: 728-730.

Bulmer M. 1991. The selection-mutation-drift theory of synonymous codon usage. Genetics 129: 897-907.

Butler G, Rasmussen MD, Lin MF, Santos MAS, Sakthikumar S, Munro CA, Rheinbay E, Grabherr M, Forche A, Reedy JL, Agrafioti I, Arnaud MB, Bates S, Brown AJP, Brunke S, Costanzo MC, Fitzpatrick DA, de Groot PWJ, Harris D, Hoyer L, Hube B, Klis FM, Kodira C, Lennard N, Logue ME, Martin R, Neiman AM, Nikolaou E, Quail MA, Quinn J, Santos MC, Schmitzberger F, Sherlock G, Shah P, Silverstein KAT, Skrzypek MS, Soll D, Staggs R, Stansfield I, Stumpf MPH, Sudbery PE, Srikantha T, Zeng Q, Berman J, Berriman M, Heitman J, Gow NAR, Lorenz MC, Birren BW, Kellis M and Cuomo CA. 2009. Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes. Nature 459: 657-662.

Clarke BC. 1970. Darwinian evolution of proteins. Science 168: 1009-1011.

Crick FH, Barnett L, Brenner S, Watts-Tobin RJ. 1961. General nature of the genetic code for proteins. Nature 192, 1227 – 1232.

Crick FH. 1966. Codon-anticodon pairing: the wobble hypothesis. J. Mol. Bio. 19:548-555.

Curran JF. 1998. Modified nucleosides in translation. In: Grosjean H, Benne R, editors. *Modification and Editing of RNA*. Washington DC: ASM Press p.493-516.

Curran JF, Yarus M. 1989. Rates of aminoacyl-tRNA selection at 29 sense codons *in vivo*. J. Mol. Biol. 209: 65-77.

Daviter T, Gromadski KB, Rodnina MV. 2006. The ribosome's response to codon-anticodon mismatches. Biochimie 88: 1001-1011.

Dong H, Nilsson L, Kurland CG. 1996. Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates. J. Mol. Biol. 260: 649-663.

Dos Reis M, Wernisch L, Savva R. 2003. Unexpected correlations between gene expression and codon usage bias from microarray data for the whole *Escherichia coli* K-12 genome. Nucleic Acids Res. 31: 6976-6985.

Dos Reis M, Savva R, Wernisch L. 2004. Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res. 32: 5036-5044.

Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on codong-sequence evolution. Cell 134, 341-342.

Duret L. 2000. tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes. Trends in Genetics. 16: 287-289.

Eyre-Walker, A. 1996. Synonymous codon bias is related to gene length in Escherichia coli: selection for translational accuracy? Mol. Biol. Evol. 13: 864-872.

Grantham R, Gautier C, Gouy M, Jacobzone M and Mercier R. 1981. Codon catalogue usage is a genome strategy for genome expressivity. Nucleic Acids Research 9: r43-r75.

Grantham R, Gautier C, Gouy M, Mercier R and Pave A, 1980. Codon catalogue usage and the genome hypothesis. Nucleic Acids Research 8: r49-r62.

Grosjean H, de Crécy-Lagard V, Marck C. 2010. Deciphering synonymous codons in the three domains of life: co-evolution with specific tRNA modification enzymes. FEBS Lett. 584(2):252-64.

Glasner JD, Liss P, Plunkett G 3rd, Darling A, Prasad T, Rusch M, Byrnes A, Gilson M, Biehl B, Blattner FR, Perna NT. 2003. ASAP, a systematic annotation package for community analysis of genomes. Nucleic Acids Res. 31: 147-151.

Gromadski KB, Daviter T, Rodnina MV. 2006. A uniform response to mismatches in codon-anticodon complexes ensures ribosomal fidelity. Molecular Cell 21: 369-377.

Gilchrist MA, Shah P, Zaretzki R. 2009. Measuring and detecting molecular adaptation in codon usage against nonsense errors during protein translation. Genetics 183(4):1493-505.

Hagervall TG, Pomerantz SC, McCloskey JA. 1998. Reduced misreading of asparagine codons by *Escherichia coli* $tRNA^{Lys}$ with hypomodified derivatives of 5-methylaminomethyl 2-thiouridine in the wobble position. J. Mol. Biol. 284: 33-42.

Henry I, Sharp PM. 2007. Predicting gene expression level from codon usage bias. Mol Biol Evol. 24(1):10-2.

Heyd A, Drew DA. 2003. A mathematical model for elongation of a peptide chain. Bull. Math. Biol. 65: 1095-1109.

Higgs PG. 2009. A Four-Column Theory for the Origin of the Genetic Code: Tracing the Evolutionary Pathways that Gave Rise to an Optimized Code. Biology Direct 4:16.

Higgs P.G. and Attwood T.K., 2005. *Bioinformatics and molecular evolution.* Blackwell Science Ltd.

Higgs PG, Jameson D, Jow H, Rattray M. 2003. The evolution of tRNA-Leucine genes in animal mitochondrial genomes. J. Mol. Evol. 57: 435-445.

Higgs PG and Ran W. 2008. Coevolution of Codon Usage and tRNA Genes Leads to Alternative Stable States of Biased Codon Usage. Mol Biol Evol. 25: 2279-91.

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

Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. 2: 13-34.

Jia WL, Higgs PG. 2008. Codon usage in mitochondrial genomes: distinguishing context-dependent mutation from translational selection. Mol. Biol. Evol. 25: 339-351.

Jühling F, Mörl M, Hartmann RK, Sprinzl M, Stadler PF, Pütz J. 2009. tRNAdb2009: compilation of tRNA sequences and tRNA genes. Nucleic Acids Res. 37: D159-D169.

Kanaya S, Yamada Y, Kudo Y, Ikemura T. 1999. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs. Gene 238: 143-155.

Kerr AR, Peden JF, Sharp PM. 1997. Systematic base composition variation around the genome of Mycoplasma genitalium, but not Mycoplasma pneumoniae. Mol. Microbiol. 25: 1177-1179.

King JL, and Jukes TH. 1969. Non-Darwinian evolution. Science 164: 788-798.

Kimura M. 1962. On the Probability of Fixation of Mutant Genes in a Population. Genetics. 1962 June; 47(6): 713–719.

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

Kothe U, Rodnina MV. 2007. Codon Reading by $tRNA^{Ala}$ with Modified Uridine in the Wobble Position. Molecular Cell, Volume 25: 167-174.

Kramer E, Farabaugh PJ. 2007. The frequency of translational misreading errors in *E. coli* is largely determined by tRNA competition. 13: 87-96.

Krüger MK, Pedersen S, Hagervall TG, Sorensen MA. 1998. The modification of the wobble base of $tRNA^{Glu}$ modulates the translation rate of glutamic acid codons in vivo, J. Mol. Biol. 284: 621-631.

Lavner Y, Kotlar D. 2005. Codon bias as a factor in regulating expression via translation rate in the human genome. Gene 345: 127-138.

Lavrov DV, Lang BF. 2005. Transfer RNA gene recruitment in mitochondrial DNA. Trends in Genetics 21: 129-133.

Lehmann J, Libchaber A. 2008. Degeneracy of the genetic code and stability of the base pair at the second position of the anticodon. RNA 14: 1264-1269.

Li WH. 1987. Models of nearly neutral mutation with particular implications for nonrandom usage of synonymous codons. J. Mol. Evol. 24: 337-345.

Lustig F, Borén T, Claesson C, Simonsson C, Barciszewska M, Lagerkvist U. 1993. The nucleotide in position 32 of the tRNA anticodon loop determines ability of anticodon UCC to discriminate among glycine codons, Proc. Natl Acad. Sci. USA 90: 3343–3347.

Lynch M, Conery JS. 2003. The origins of genome complexity. Science 302: 1401-1404.

Meier F, Suter B, Grosjean H, Keith G, Kubli E. 1985. Queuosine modification of the wobble base in $tRNA^{His}$ influences *in vivo* decoding properties. EMBO J. 4: 823-827.

Mitarai N, Sneppen K, Pedersen S. 2008. Ribosome collisions and translational efficiency: optimization by codon usage and mRNA destabilization. J. Mol. Biol. 382: 236-245.

Morris RC, Brown KG, Elliott MS. 1999. The effect of queuosine on tRNA structure and function. J Biomol Struct Dyn 16: 757–774.

Morris RC, Elliott MS. 2001. Queuosine modification of tRNA: A case for convergent evolution. Mol. Genet. Metab. 74: 147-159.

Nakao A, Yoshihama M, Kenmochi N. 2004. RPG: the Ribosomal Protein Gene Database. Nucleic Acids Res. 32: D168-D170.

Näsvall SJ, Chen P, Björk R. 2007. The wobble hypothesis revisited: uridine-5-oxyacetic acid is critical for reading of G-ending codons. RNA 13: 2151–2164.

Ninio J. 2006. Multiple stages in codon-anticodon recognition: double-trigger mechanisms and geometric constraints. Biochimie 88: 963-992.

Pearson, K. 1901. On Lines and Planes of Closest Fit to Systems of Points in Space. Philosophical Magazine 2 (6): 559–572.

Phelps SS, Malkiewicz A, Agris PF, Joseph S. 2004. Modified nucleotides in $tRNA^{Lys}$ and $tRBA^{Val}$ are important for translocation. J. Mol. Biol. 338: 439-444.

Precup J, Parker J. 1987. Missense misreading of asparagine codons as a function of codon identity and context. J. Biol. Chem. 23: 11351-11355.

Percudani R, Pavesi A, Ottonello S. 1997. Transfer RNA gene redundancy and translational selection in *Saccharomyces cerevisiae*. J. Mol. Biol. 268: 322-330.

Ran W and Higgs PG. 2010. The influence of anticodon-codon interactions and modified bases on codon usage bias in bacteria. Mol Biol Evol. doi:10.1093/molbev/msq102

Rocha EPC. 2004. Codon usage bias from the tRNA's point of view: Redundancy, specialization, and efficient decoding for translational optimization. Genome Res. 14: 2279-2286.

Rodnina MV, Wintermeyer W. 2001. Fidelity of aminoacyl-tRNA selection on the ribosome: Kinetic and structural mechanisms. Annu. Rev. Biochem. 70: 415-435.

Romier C, Ficner R, Suck D. 1998. Structural basis of base exchange by tRNA-guanine transglycosylases. In: Grosjean H, Benne R, editors. *Modification and Editing of RNA*. Washington DC: ASM Press p.493-516.

Saks ME, Sampson JR, Abelson J. 1998. Evolution of a transfer RNA gene through a point mutation in the anticodon. Science 279: 1665-1667.

Selinger DW, Cheung KJ, Mei R, Johansson EM, Richmond CS, Blattner FR, Lockhart DJ, Church GM. 2000. RNA expression analysis using a 30 base pair resolution *Escherichia coli* genome array. Nat. Biotechnol.18: 1262-8.

Sengupta S, Higgs PG. 2005. A Unified Model of Codon Reassignment in Alternative Genetic Codes. Genetics 170: 831-840.

Sengupta S, Yang X, Higgs PG. 2007. The mechanisms of codon reassignments in mitochondrial genetic codes. J. Mol. Evol. 64: 662-688.

Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE. 2005. Variation in the strength of selected codon usage bias among bacteria. Nucleic Acids Res. 33: 1141-1153.

Sharp PM, Cowe E, Higgins DG, Shields DC, Wolfe KH, Wright F. 1988. Codon usage patterns in *Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster* and *Homo sapiens*; a review of the considerable within species diversity. Nucleic Acids Res. 16: 8207-8211.

Sharp PM, Li WH. 1986. Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for 'rare' codons. Nucl. Acids Res. 14: 7737-7749.

Sharp PM, Li WH. 1987. The codon adaptation index – a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15: 1281-1295.

Solomovici J, Lesnik T, Reiss C 1997. Does *Escherichia coli* optimize the economics of the translation process? J. Theor. Biol. 185: 511-521.

Sorensen MA, Pedersen S. 1991. Absolute *in vivo* translation rates of individual codons in *Escherichia coli*. J. Mol. Biol. 222: 265-280.

Sprinzl M, Vassilenko K. 2005. Compilation of tRNA sequences and sequences of tRNA genes. Nucleic Acids Res. 33:D139-40.

Stoletzki N, Eyre-Walker A. 2006. Synonymous codon usage in *Escherichia coli*: Selection for translational accuracy. Mol. Biol. Evol. 24: 374-381.

Urbonavicius J, Qian Q, Durand JMB, Hagervall TG, Björk GR. 2001. Improvement of reading frame maintenance is a common function for several tRNA modifications. EMBO J. 20: 4863-4873.

Vendeix FAB, Dziergowska A, Gustilo EM, Graham WD, Sproat B, Malkiewicz A, Agris PF. 2008. Wobble-position modifications contribute order to tRNA's anticodon for ribosome-mediated codon binding. Biochemistry 47: 6117–6129.

Watson JD. and Crick. FH 1953. A structure for deoxyribose nucleic acids. *Nature* 171:737-738.

Weixlbaumer A, Murphy FV, Dziergowska A, Malkiewicz A, Vendeix FAP, Agris PF, Ramakrishnan V. 2007. Mechanism for expanding the decoding capacity of transfer RNAs by modification of uridines. Nat Struct Mol Biol 14: 498–502.

Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol. 18(5):691-9.

Withers M, Wernisch L, dos Reis M. 2006. Archaeology and evolution of transfer RNA genes in the *Escherichia coli* genome. RNA 12: 933-942.

Wright F. 1990. The effective number of codons used in a gene. Gene 87: 23-29.

Xia X. 2008. The cost of wobble translation in fungal mitochondrial genomes: integration of two traditional hypotheses. Submitted to BMC Evolutionary Biology.

Yarian C, Townsend H, Czestkowski W, Sochacka E, Malkiewicz AJ, Guenther R, Miskiewicz A, Agris PF. 2002. Accurate translation of the genetic code depends on tRNA modified nucleosides. J. Biol. Chem. 277: 16391–16395.

Yokobori S, Suzuki T, Watanabe K. 2001. Genetic code variantions in mitochondria: tRNA as a major determinant of genetic code plasticity. J. Mol. Evol. 53: 314-326.

Yokoyama S, Watanabe T, Murao K, Ishikura H, Yamaizumi Z, Nishimura S, Miyazawa T. 1985. Molecular mechanism of codon recognition by tRNA species with modified uridine in the first position of the anticodon. Proc. Nat. Acad. Sci. USA. 82: 4905-4909.

Zouridis H, Hatzimanikatis V. 2007. A model for protein translation: Polysome self-organization leads to maximum protein synthesis rates. Biophys. J. 92: 717-730.

Zouridis H, Hatzimanikatis V. 2008. Effects of codon distributions and tRNA competition on protein translation. Biophys. J. 95: 1018-1033.