School of Crystallography

Birkbeck College

University of London

Understanding the Patterns of Codon Usage Bias in Prokaryotic and Eukaryotic Genomes

Mario José dos Reis Barros

Work conducted under the supervision of

Professor Lorenz Wernisch

and submitted for the degree of

Doctor of Philosophy of The University of London

This manuscript has been entirely written by the author. Unless otherwise explicitly stated, all the results and analysis presented herein have been carried out by the author.

Abstract

The genetic code is redundant. This means that most amino acids are encoded by several codons. Different synonymous codons are used to different extents and every genome shows particular codon preferences. Despite nearly three decades of research, how codon preferences come about remains an enigmatic riddle in molecular evolution. It is generally acknowledged that optimal codons help achieve translational optimisation and accuracy, and they reflect the anticodon composition of the genomic tRNA pool. The first part of the riddle then is why there should be such wide differences in genomic tRNA pool size and anticodon composition. Furthermore, codon usage-tRNA optimisation is only seen in a fraction of prokaryotic and eukaryotic genomes, so the second part of the riddle is why natural selection should act on certain genomes to optimise codon usage while ignoring others. This work explores patterns of codon usage in prokaryotic and eukaryotic genomes. A method is developed to test whether natural selection is active at the synonymous codon level, and the method is applied to several genomes. It is shown that selection on codon usage tends to be maximal for organisms with intermediate genome size. Ancestral reconstruction of tRNA sets and codon sequences are carried out to understand the evolution of tRNA genes in bacterial genomes and whether there is correlation between tRNA and codon usage evolution. In organisms were selection is operative, the particular codon preferences might be determined by the evolutionary history of their genomic tRNA pool. The work ends with a general discussion of how the reduction in population size associated with the evolution of genome complexity and the stochastic processes involved in tRNA evolution might explain if natural selection is operative on silent sites in different genomes.

Abbreviations

A,T,U,C,G	Adenine, Thiamine, Uracil, Cytosine, Guanine
CAI	Codon adaptation index
DNA	Dexoxyribonucleic acid
*Ec*K12	*Escherichia coli* K-12
*Ec*O157	*Escherichia coli* O157:H7 EDL933
*Ec*CFT073	*Escherichia coli* CFT073
GC3s	G + C content at third synonymous position in codons
GP	Gaussian process
CpG	Cytosine-phosphate-Guanine
K2P	Kimura's two parameter model
MCMC	Markov Chain Monte Carlo
ML	Maximum likelihood
mRNA	Messenger RNA
Myr	Million years
Nc	Effective number of codons
ORF	Open reading frame
PIC	Phylogenetically independent contrasts
RNA	Ribonucleic acid
rRNA	Ribosomal RNA
SF1, SF2	Synonymous family 1, synonymous family 2, etc.
*Sf*2a301	*Shigella flexneri* 2a301

*St*LT2	*Salmonella typhimurium* LT2
tAI	tRNA adaptation index
tRNA	Transfer RNA
UPGMA	Unweighted pair group with arithmetic mean

Preface

The genetic code is redundant, and this redundancy means that most amino acids are coded for by more than one codon. Before the first DNA sequences became available, it seemed natural to think that the codons that were used to encode a particular aminoacid in a particular gene and in a particular organism would be drawn randomly from the available set of codons for that amino acid. It was surprising then when the first DNA sequences were obtained in the late '70s, and it was observed that certain codons were consistently being chosen to encode particular amino acids. A lot of excitement and research followed, and it was quickly shown that in fast growing microorganisms such as *Escherichia coli* or the baker's yeast the preferred codons were those recognised by the most abundant cognate transfer RNAs within the cell. This was interpreted as the signature of natural selection acting at the silent DNA level to optimise protein translation within the cell. Although initially this seemed to contradict the main postulate of the neutral theory, that most evolution at the molecular level is driven by the random fixation of mutants that are selectively neutral, further work has since reconciled the action of selection on silent sites with this postulate. Much research has been carried out since those first observations, and it has been shown that natural selection is quite pervasive in optimising codon usage in disparate looking organisms such as yeasts, flies, worms and certain bacteria, while seemingly ignoring others such as humans, mice and other bacteria. Although some patterns have emerged, despite nearly three decades of research, there is still a lot of confusion surrounding the idiosyncratic behaviour of

natural selection at silent sites.

This works summarises my personal three-year attempt at finding some answers to the questions presented above. This work is organised in the following manner: The first chapter is an introduction to the genetic code, its codons, and the riddles of codon usage bias. The second chapter provides an introduction to molecular evolutionary genetics and it discusses the population genetic models that are necessary to study and understand the problem of selection at codon sites. The third chapter deals with measuring codon usage, and some classical indexes to measure codon usage are discussed. Here I introduce a new index that takes into account the tRNA composition of a given genome, and a test is developed to assess whether codon usage in such genome has been shaped by natural selection. In chapter four the ideas developed in chapter three are applied to the analysis of selection and codon usage in a large number of eukaryotic and prokaryotic genomes, with some interesting findings. In chapter five I explore the seemingly ignored issue of tRNA evolution and its influence on the evolution of codon usage, using *Escherichia coli* as a model organism. Chapter six deals with ancestral codon reconstruction in *Escherichia coli* with some surprising results. Finally, chapter seven presents a final exploratory analysis on selection, codon usage and population size in eukaryotic organisms, and the results from the previous chapters are summarised. A general model of codon usage is proposed, stressing future lines of research in this area.

Parts of this manuscript have already been published in a more succinct and modified form. Part of the contents of chapters 3 and 4 appeared in

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

chapter 5 in

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

and chapter 7 in

> dos Reis M. and Wernisch L. (2009) Estimating translational selection in eukaryotic genomes. Mol. Biol. Evol., 26:451-461.

I am indebted to Professor Lorenz Wernisch for his help and supervision during the past three years. I would also like to thank him for having granted me the freedom to develop my own research path, and for having fought fiercely in order to secure the studentship grant that allowed me to pursue this work. I would also like to thank Mike Withers for his contribution on the the analysis of tRNA evolution in *Escherichia coli* (chapter 5). Mike took this task as part of his MSc Bioinformatics project, which I had the pleasure to supervise. He identified and compiled tRNA sequences for various *E. coli* genomes, and wrote several useful scripts to calculate pairwise genetic distances between different tRNA genes. He constructed genetic distance matrices for all the tRNA genes that were analysed. These matrices were the basis for the construction of the synthenic map of tRNA evolution in *E. coli* presented in chapter 5. I am also indebted to several people that at some point shared their scientific views with me: Michael Lynch for showing me his unpublished chapter on the nucleotide composition landscape of genomes, and for sharing his ideas on the role of genome complexity on codon usage evolution; Paul Sharp for his constructive criticism on my test of tRNA-codon coadaptation in genomes; Eduardo Rocha for comments on his tRNA paper; Adam Eyre-Walker, Ziheng Yang, Michael Bulmer and Renos Savva for several useful comments. This work was supported by a studentship granted by the Biotechnology and Biological Sciences Research Council.

Mario dos Reis

London

February, 2009

Contents

List of Tables

List of Figures

1 Riddles of Codon Usage

1.1 A brief introduction to the genetic code

The fifties and sixties saw a revolution in our understanding of evolution at the molecular level. Two biological breakthroughs took place during those decades. First, Watson and Crick [141] elucidated the structure of the Dexonyribonucleic Acid (DNA) molecule. This provided unprecedented insight into how the genetic information was stored and how it could be replicated within organisms. Watson and Crick showed that DNA was formed by two strands of nucleotides twisted together forming a double helix, with the different nucleotidic bases neatly stacked, the bases on one strand forming specific hydrogen bonds with the ones on the opposite strand: Adenine (A) with Thiamine (T), and Cytosine (C) with Guanine (G). Because both strands are complementary, this immediately suggested a mechanism of replication: each strand could serve as the template onto which two new strands could be built, forming two DNA molecules identical to the parental one. The second breakthrough was the elucidation of the genetic code. What seemed like a mere dream to molecular biologists at the beginning of the decade became a reality in a handful of years. The information contained in DNA is first transcribed into an intermediate Ribonucleic Acid (RNA) molecule. This messenger molecule then couples with the ribosomes and transfer RNAs and protein synthesis takes place. Proteins are made up of combinations of twenty possible amino acids. The exact sequence of amino acids that made up a particular protein is determined by the

nucleotide sequence of the DNA molecule. Because DNA is composed of only four nucleotides, combinations of at least three nucleotides are needed to encode all twenty amino acids. There are $4^3 = 64$ possible permutations of four nucleotides. Each permutation, is a word, or codon that translates exactly into one unique amino acid. The elucidation of the meaning of these 64 codons, the discovery of the genetic code, is one of the great scientific achievements of the twentieth century. For an historic perspective on these discoveries the reader is referred to [109].

An important feature of the genetic code is that it is nearly universal. Organisms as different as the bacterium *Escherichia coli*, which lives in the intestines of mammals, the baker's yeast *Saccharomyces cerevisiae*, or the common farm chicken *Gallus gallus*, all use the same code to translate the genetic information contained in their genomic DNA to manufacture the proteins that form the chemical machinery of the cell. It is common practise in molecular biology to introduce human DNA into *E. coli* cells, where the information contained in the DNA molecule is correctly translated by the bacterium machinery in order to produce the original human proteins. If the genetic code were not universal, this would not be possible. Exceptions to the universality of the genetic code exist. For example, cellular organelles like mitochondria or chloroplasts have their own DNA chromosomes, and the genes contained within them are translated under slightly modified variations of the code. The variations in the genetic code are reviewed in [109, 67]. For convenience, the nearly universal code is simply called the standard code.

Because there are 64 words to encode only 21 meanings (the twenty amino acids plus the translation termination signal), the genetic code is said to be degenerate or redundant. Codons that encode the same amino acid are called synonymous, and those that encode different amino acids are said to be non synonymous. The redundancy in the code has been shown to give it resilience against mutations [43]. Because of the code redundancy, a large proportion of substitution mutations only change one codon by a synonymous one. In fact most mutations at the third position

(or site) in a codon are synonymous. Some mutations at the first position are also synonymous, while mutations in the second position are always non synonymous. Since amino acids with similar physicochemical properties are neighbours in the code, most non synonymous mutations change one amino acid by a similar one, minimising any possible impact on the structure of the corresponding protein. The genetic code has been shown to be highly optimised to minimise the effects of deleterious mutations at the DNA level [43].

The twenty amino acids can be classified according to how many synonymous codons encode each one of them. So for example, the amino acid Methionine is encoded by only one codon, and so is Tryptophan. These two amino acids are said to belong to the synonymous family one (SF1). There are nine amino acids encoded by two codons (SF2); only one encoded by three codons (SF3); five encoded by four codons (SF4) and three encoded by six codons (SF6). Table 1.1 shows the standard genetic code and the synonymous families.

1.2 A Standard language with many dialects

As we have seen, the degeneracy of the genetic code means that most amino acids are coded for by more than one codon. Before the first gene sequences became available, it was natural to assume that the codon that would be used to code for a particular amino acid in a particular gene would be drawn randomly from the set of available codons for that amino acid. When the first genes were sequenced in the late 70s, it was noticed that the use of synonymous codons in some of these genes departed from randomness. This was specially evident in the ribosomal protein genes of *Escherichia coli* [116, 115]. Post *et al*. [115] proposed that the codons preferentially used were those efficiently recognised by the most abundant tRNA species, an idea that had previously been suggested by Clarke [20], and that was later experimentally confirmed by Ikemura [61]. A lot of work and excitement

Table 1.1: The standard genetic code.

1	2				3
	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
U	Phe	Ser	Tyr	Cys	C
U	Leu	Ser	Stop	Stop	A
U	Leu	Ser	Stop	Trp	G
C	Leu	Pro	His	Arg	U
C	Leu	Pro	His	Arg	C
C	Leu	Pro	Gln	Arg	A
C	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
A	Ile	Thr	Asn	Ser	C
A	Ile	Thr	Lys	Arg	A
A	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
G	Val	Ala	Asp	Gly	C
G	Val	Ala	Glu	Gly	A
G	Val	Ala	Glu	Gly	G

SF1 SF2 SF3 SF4 SF6

Colo rs indicate the different synonymous families. Synonymous Family 1 (SF1), synonymous family 2 (SF2), etc. Ala, Alanine (A); Arg, Arginine (R); Asn, Asparagine (N); Asp, Aspartic acid (D); Cys, Cysteine (C); Gln, Glutamine (Q); Glu, Glutamic acid (E); Gly, Glycine (G); His, Histidine (H); Ile, Isoleucine (I); Leu, Leucine (L); Lys, Lysine (K); Met, Methionine (M); Phe, Phenylalanine (F); Pro, Proline (P); Ser, Serine (S); Thr, Threonine (T); Trp, Tryptophan (W); Tyr, Tyrosine (Y); Val, Valine (V); Stop, translation termination codon. In the mRNA molecule, Uracil (U) substitutes Thiamine, so the standard code is usually represented in terms of U rather than T.

followed in this area of research. Grantham *et al*. [49] provided the first compilation of codon usage trends for various viral, bacterial, organelle and Eukaryotic genomes. These workers used a multivariate analysis technique known as correspondence analysis to project the codon data from a high dimensional space into a two dimensional plane for easy visualisation. They repeated the analysis on amino acids, and showed that while different genomes tend to use very different codons, they show similar amino acid usage tendencies (figure 1.1). These led to the postulation of the genome hypothesis which states that the genes in a particular genome conform to its characteristic codon usage catalogue [50]. So although the genetic code is nearly universal, different genomes seem to speak different dialects. This is easily seen when human DNA is introduced into *E. coli* cells. Although the human genes are correctly translated into the corresponding proteins, the translation process itself is usually inefficient. If the codons in the human genes are modified to resemble those of *E. coli*, much better expression of the human proteins can be achieved (e.g. [51]). Another good example is the green fluorescent protein, which is commonly used in Eukaryotic cell systems for the study of gene expression. This protein was originally obtained from a jellyfish (*Aequorea victoria*), but the gene normally used in mammalian cell transfection experiments, usually uses a codon optimised version of the original gene that has better expression in the mammalian host cells [88].

1.3 Riddles of codon usage

Much work has been carried out on codon usage research since the phenomenon was first noticed in the late 70's. In the early 80's, studies in the bacterium *E. coli* and in the baker's yeast *Saccharomyces cerevisiae* showed that highly expressed genes tend to use a subset of optimal codons that are recognised by the most abundant cognate tRNAs within the cell [61, 10]. The interesting thing being that some op-

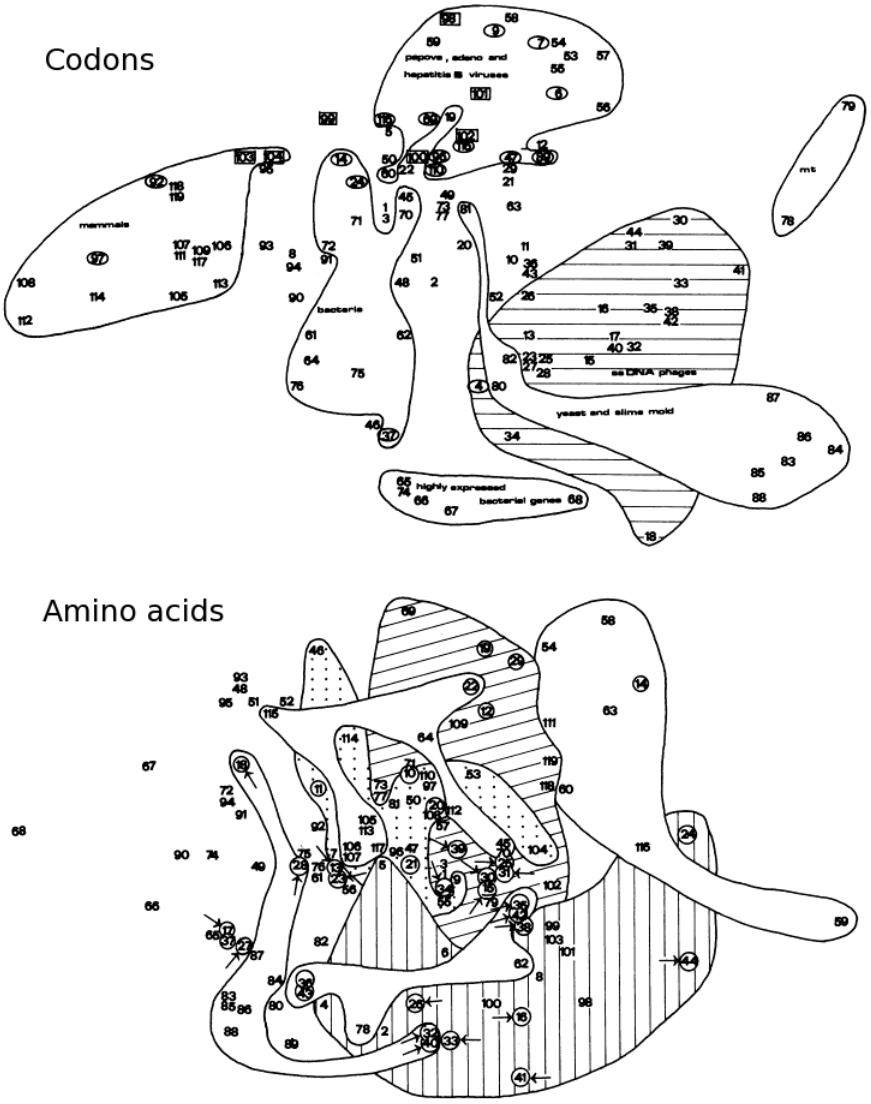

Figure 1.1: Multivariate analysis on codon and amino acid usage for several genomes.

A set of n coding sequences from different organisms can be arranged in a n x 59 table where the columns are the codon counts in the sequence (Stop, Met and Thr codons are usually ignored). Each gene is then a point in a 59 dimensional space. Genes that cluster together in this space show similar codon usage trends. Visualising data in high dimensional spaces is an arduous task for human beings. For this reason computers are usually used to project the data from higher dimensional spaces into a few dimensions (usually two or three) so it can be more easily visualised. In the technique of correspondence analysis, the absolute codon frequencies are first scaled appropriately and linear algebra is used to find the axes that explain the largest variation in the data. Correspondence analysis has been quite popular in codon bias research (see for example [102, 113]). Reprinted from Grantham *et al.* [49].

timal codons in yeast were different from those in *E. coli*, reflecting the underlying differences in the tRNA composition between both genomes. The form of natural selection that acts at the level of codon usage optimisation was coined as translational selection. These first studies hypothesised that the use of optimal codons in highly expressed genes would help achieve faster translation rates in the ribosomes and also a higher translation accuracy [8]. This led to the idea that translational selection was a feature of unicellular organisms. Studies in the late 80's and early 90's showed first in the common fruit fly (*Drosophila melanogaster*) [130, 2] and later on a nematode worm (*Caenorhabditis elegans*) [131] that synonymous codon usage in multicellular organisms could also be subject to the action of natural selection. However, early studies in mammals showed no evidence of natural selection acting at the level of synonymous codon usage [41]. Furthermore, some studies also showed a lack of selected codon usage in certain unicellular microorganisms [83]. At this point a general picture was emerging where translational selection was pervasive, and present in a wide variety of (but not all) organisms. Furthermore, selection on synonymous codon usage seemed to vary in strength within (*i.e.* highly vs lowly expressed genes) and among genomes (*e.g.* fast growing microorganisms vs multicellular Eukaryotes). A mutation selection balance theory of synonymous codon usage emerged [13, 125] that explains patterns of codon usage in different organisms as a balance between the intrinsic mutational bias of a given genome, and the effect of translational selection on highly expressed genes.

There is still debate on whether the chief cause of selected codon usage is due to selective pressure for translational speed or pressure to reduce mistranslational errors. Some of the latest research indicates that translational accuracy might be the most important factor (see for example [132] and [31]). Because translational errors could lead to the accumulation of toxic, misfolded proteins, selective pressure against such errors would be expected to be substantial, even for large multicellular organisms [31]. Indeed, recent work indicates that for the most highly expressed

genes in mammals, selected codon usage is weak, but present nonetheless [137, 23, 153, 31].

As we have seen, there is a wide range of codon usage patterns across living organisms. And those patterns were explained by different mutational biases characteristic to each genome, and by whether translational selection was operative or not. However the situation has been quite confusing. First, why are there such wide differences in genomic tRNA pool size and anticodon composition in Eukaryotes and Prokaryotes? Furthermore, codon usage-tRNA optimisation is only seen in a fraction of all the genomes analysed, so why should natural selection act on certain genomes to optimise codon usage while ignoring others (table 1.2)? Despite nearly three decades of research on codon usage, this questions have not been satisfactorily answered [135].

Before we set out to try and answer the enigmatic questions presented above, some background in molecular evolution and population genetics is necessary. That is the topic of the next chapter. In later chapters we will explore the problem of measuring synonymous codon usage bias and the difficulties in estimating translational selection. Once these hurdles have been cleared, we will explore the patterns of codon usage and translational selection from Prokaryotes to Eukaryotes with the hope of providing some more light onto the issues presented above.

Table 1.2: Selection on codon usage for some representative genomes

Super Kingdom	Organism	Selection?	References[1]
Prokaryota	*Escherichia coli*	yes, strong	[60]
	Bacillus subtilis	yes, weak	[69, 103]
	Helicobacter pylori	no	[83]
	Borrelia burgdorferi	yes, no	[99, 84, 113]
Eukaryota	Baker's yeast	yes, strong	[10]
	Drosophila melanogaster	yes, moderate	[3]
	Caenorhabditis elegans	yes, moderate	[131]
	Arabidopsis thaliana	yes, weak	[34]
	Homo sapiens	weak, no	[70, 23, 137, 86, 153]

[1]The literature on codon usage bias is quite large. The references provided here do not aim to be complete or extensive.

2 Codon Usage and Molecular Evolution

2.1 The importance of chance in evolution

If we throw a fair coin ten times, we would expect to see five tails and five heads. However we know, by practical experience, that we might as well see six tails and four heads, seven heads and three tails or some other result. This is simply a natural consequence of the fact that the system is stochastic. The expected value is simply the average outcome when the experiment is repeated a very large number of times, but some times the expected value is not even a possible outcome (if we throw the coin seven times, we would expect to see 3.5 heads, but this is impossible!). Let us imagine a population of N diploid organisms. Each individual has two chromosomes so there are $2N$ chromosomes in the population. In the population there is a gene a that determines eye colour, let us write a_1 for the allele type that induces blue eyes, and a_2 for the type that induces green eyes. We are not interested in knowing whether a_1 is dominant over a_2 or not, in the following discussion this is irrelevant. There are exactly N copies of a_1 in the population (and hence N copies of a_2). Let us also imagine that these organisms are seasonal, they mate randomly at the beginning of the season, lay exactly N eggs, and the adults die. The next season the eggs hatch, the newly grown adults reproduce, and the process is repeated all over again, season after season, generation after generation. The question is, what happens to

alleles a_1 and a_2 throughout time? Let us focus on the first generation. When the adults mate, they will provide $2N$ copies of gene a to their offspring. This is similar to the coin example above. We know the expected number of copies of a_1 that will be passed to the next generation is N. However, the actual number of copies that will be passed is a random variable that depends on the stochastic nature of mating and gamete sampling to produce new individuals. We can see that throughout generations the actual number of copies of a_1 will fluctuate randomly. This fluctuation of allele frequencies throughout generations is what is known as random genetic drift. Now let us imagine that $N = 2$. The four alleles present in the first generation will be $a_1a_1a_2a_2$. After gamete sampling, the next generation could be composed like $a_1a_2a_2a_2$ or $a_1a_1a_1a_2$ or one extreme case such as $a_1a_1a_1a_1$. There are in fact, 16 possible outcomes. It is easy to see that as the frequency of a_1 fluctuates, a_1 will eventually be fixed or will be lost altogether from the population. Once the population reaches either one of these two states, it becomes stuck, and only mutation or migration can reintroduce genetic variation into the population.

So an interesting question emerges. What is the probability that allele a_1 will become fixed after $n \to \infty$ generations? As it turns out, the answer is very simple (although the mathematical proof is complicated [72]). The probability of fixation of allele a_1 is just its initial frequency. In the example above, the initial frequency would be $p_0 = N/2N = 0.5$. So there is a 50% chance that all the organisms in our imaginary population will have blue eyes (and symmetrically, green eyes) after a sufficiently large number of generations have taken place. This elegant result has surprising implications. Let us imagine the following scenario: a certain diploid population of grey hairy animals has one million individuals; one single calf is born that carries a dominant mutation that makes its fur black. This little calf is neither fitter nor weaker than the rest of the individuals in the population. Hence, the black allele is said to be selectively neutral. Is there any chance that the black fur condition will ever spread throughout the population? Well, the answer is yes. The

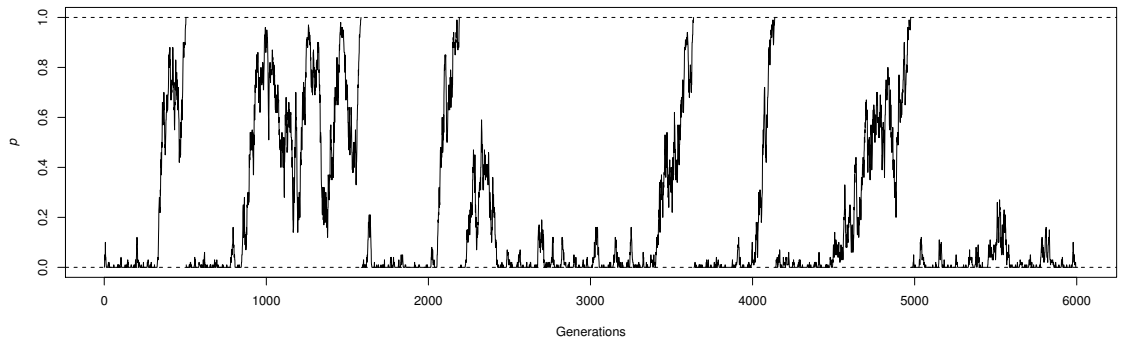

Figure 2.1: Fixation of neutral alleles
A simple computer simulation of two neutral alleles with reversible mutation. It can be seen that although the initial frequency of a new mutant is very low, it can, sometimes, overtake the population and reach fixation. Most mutations are lost (the small peaks at the bottom), while very few reach fixation (six in this case, top of the figure).

black allele has an initial frequency $p_0 = 1/2N$, so it has a probability of one in two million of becoming fixed in the population! This might seem like a very remote chance, and indeed it is, but when we take into account that new different mutants appear continuously, generation after generation, then, at some point, one lucky mutant allele will eventually spread from one single copy to the whole population. This is one of the most important results of the stochastic theory of population genetics.

We can take the argument even further, and notice that a neutral mutant allele that has reached fixation in a population has only a limited lifespan, since at some point in the future it will be substituted by another new mutant (figure 2.1). Those new neutral mutants that are destined to become fixed in a diploid population take an average of $4N$ generations to reach fixation. The fate of most mutant alleles is to be lost, since their initial frequencies are very low, and hence have a very small probability of fixation. Another important result from stochastic population genetics is that the rate of fixation of neutral mutants is the same as the mutation rate, and it is independent of the population size.

A useful way to think about random drift is to imagine a series of independent

populations that evolve simultaneously. Then we can ask what the average behaviour of the populations would be. For example, in our original example of green and blue eyed organisms, we can imagine a large population of size $10N$ that becomes split into 10 smaller isolated sub populations (such as a series of 10 small islands). Each new sub population has size N, and considering that there is no migration between islands, the evolution of each sub population will be independent from the other ones. We can easily see that after thousands of years of isolation, the sub populations will become monomorphic, some islands (perhaps five) will contain only blue eyed individuals, while the remaining ones will contain only green eyed individuals.

So far we have considered the case of alleles that are selectively neutral. However we should also consider the interesting case of natural selection under random drift. Let us imagine that the blue eyed individuals can see far better, can find more food, and hence their own survival, reproductive rate, and the survival of their youngsters is better than in the green eyed counterparts. Deterministic models of evolution tell us that if we start with an equal number of blue and green eyed individuals, eventually, the fit blue eyed ones will overtake the whole population. But would the same happen if random drift is taken into account? Well it depends. Random drift introduces noise into the evolutionary process. However this noise depends on the population size. Generation after generation, allele frequencies will change by small steps in large populations, and by larger steps in small ones. The consequence of this is that natural selection is more effective the larger the population, because the system tends to behave in a more deterministic manner. In population genetics selection is usually represented as the selection coefficient s. In our example, let us assume that ten blue eyed individuals survive each season for every eight green eyed. We said that the relative fitness of a_1 is 1, and the fitness of a_2 is $1-s=0.8$, where s, the selective coefficient against a_2, is 0.2. Models of random drift usually assume ideal populations with special properties (such as constant size, equal

number of males and females, etc.). Real populations are not ideal, so usually an effective population size N_e is computed, and used in the mathematical models rather than the actual population size N. A real population of size N will suffer the same degree of random drift as the corresponding ideal population of size N_e. Stochastic population models of selection and drift tell us that if the product $|N_e s| \ll 1$, then the fate of an advantageous (or disadvantageous) allele would be similar to that of a neutral allele [76]. In fact, selection modifies the probability of fixation of an allele under random drift as a function of population size (figure 2.2, [72]). In small populations alleles behave as if they were neutral, their probability of fixation being the same as their initial frequencies. For very large populations the probability of fixation is practically one. As a corollary, new mutants that only provide a small selective advantage to their carrier will only be successful if this mutations happen to appear in a large population (figures 2.3 and figure 2.4).

In 1968 Kimura [73] published a paper stating that the rate of evolution at the molecular level could be better understood under a neutral drift model. He proposed that most new mutants in a population are either highly deleterious, in which case they never become fixed, or selectively neutral, in which case their fate is largely determined by random drift. The evolutionary rate for a protein would then be averaged among the fraction of conserved sites (which evolve with rate zero) and the variable neutral sites (that evolve at the neutral mutation rate). A year later, King and Jukes [78] published a paper stating that a lot of evolution at the protein level was non-Darwinian, *i.e.* independent of the action of natural selection. Jukes had been working with cytochrome c proteins, and it had seemed to him that there was too much variation in the amino acid composition of this protein across different species, to the extent that it seemed unnecessary to its function [65]. The works by Kimura and King and Jukes were the birth of the neutral theory of molecular evolution. The main postulate of the neutral theory is that most evolution at the molecular level happens without the action of natural selection. We will not discuss

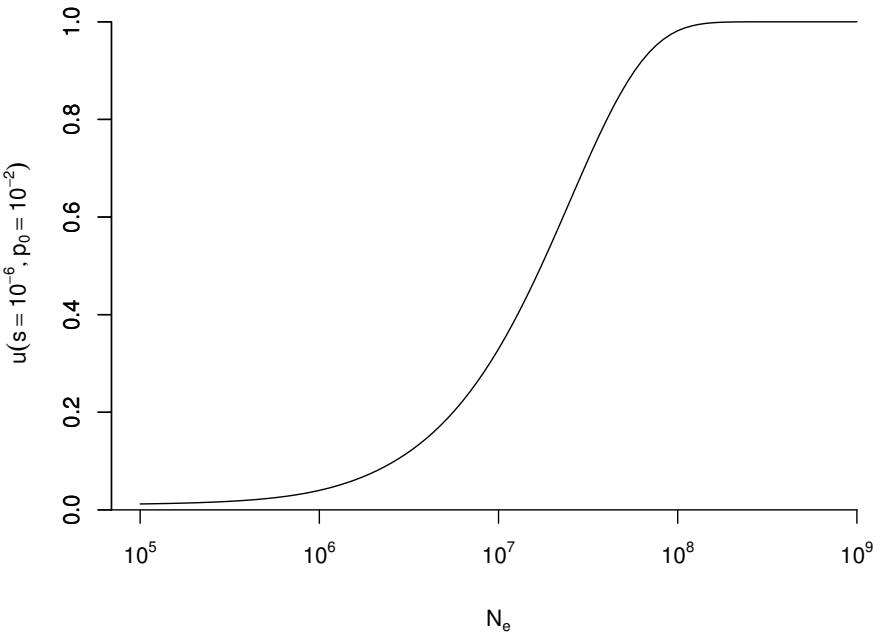

Figure 2.2: Probability of fixation of an advantageous allele as a function of the effective population size.
For very small populations, the probability of fixation is simply the initial frequency of the allele (p_0). As population size increases, the probability of fixation approaches one, and the system goes from a stochastic to a deterministic behaviour.

here the merits or drawbacks of this postulate. The important thing is that the mathematical theory developed to support the neutral theory is one of most elegant ones in biology. The equations describing the fate of neutral mutants in a population, and the effects of random drift on selected mutants are seldom disputed. One of the most important contributions of the neutral theory to modern evolutionary thinking has been its stress on the role of chance in evolution. For an authoritative source of stochastic process in population genetics the reader is advised to consult the book by Crow and Kimura [25]. The neutral theory itself is extensively discussed in another book by Kimura [76]. This later work is highly recommended.

The stochastic nature of evolution in small populations has important implications for the understanding of codon usage evolution. Selection coefficients for translational optimisation are relatively small [54], so we expect translational se-

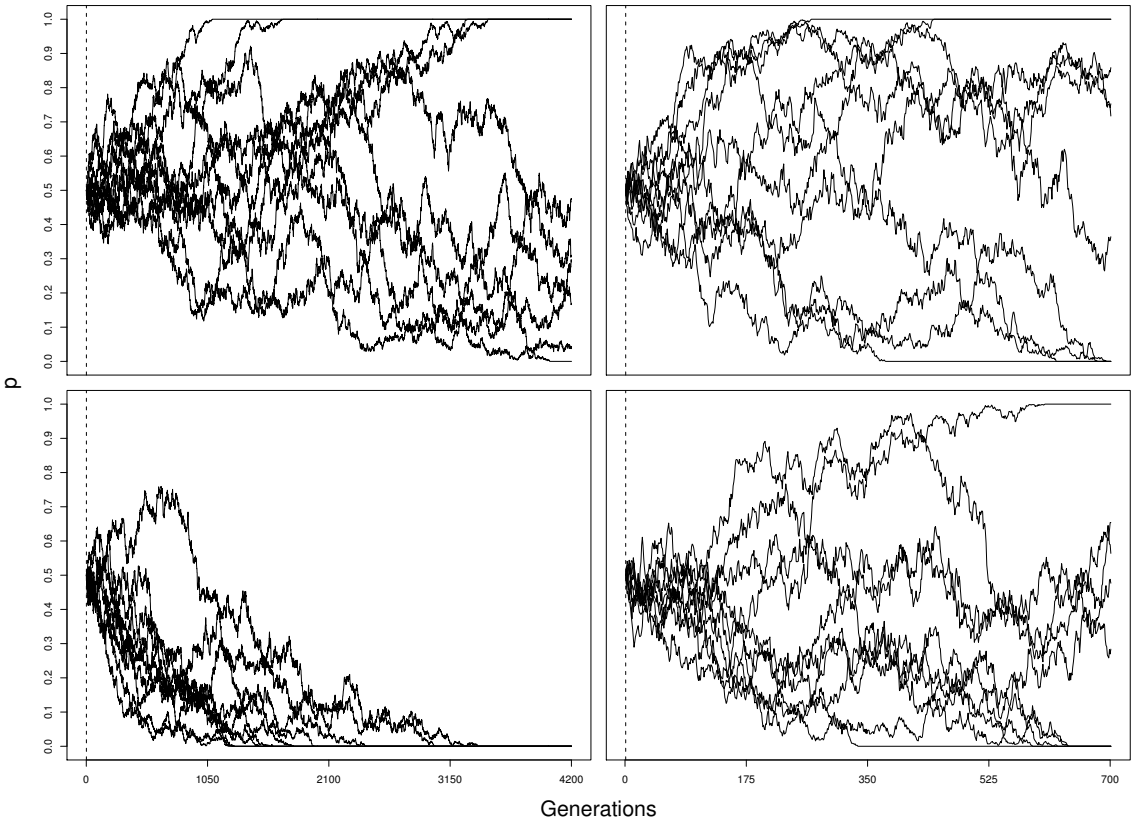

Figure 2.3: Random drift vs natural selection
Computer simulations of random drift in large (left, $N = 4,200$) and small (right, $N = 700$) populations, under no selection (top, $s = 0$) and under weak deleterious selection (bottom, $s = 2 \times 10^{-3}$). It can be seen that weak selection is very effective in driving allele frequency change in large populations, but its effect is all but negligible in small ones. Each line represents the evolutionary fate of one single population.

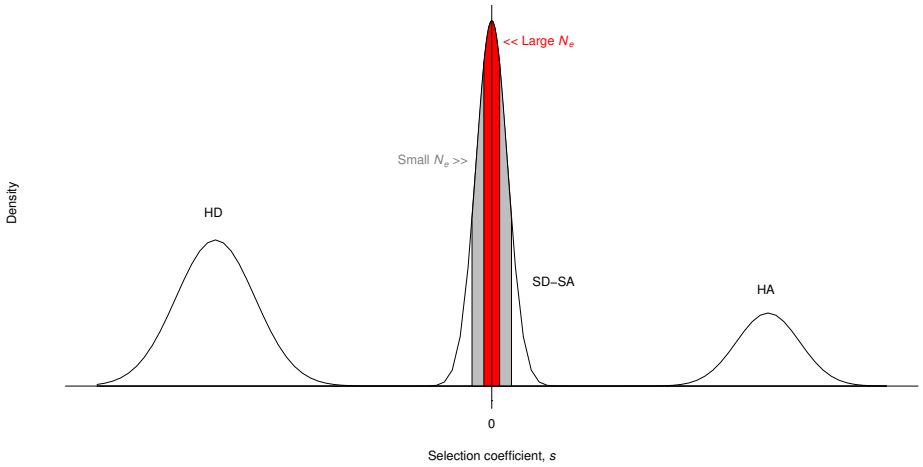

Figure 2.4: Distribution of fitness effects of novel mutants in a population. When a new mutant appears in a population it can be Highly deleterious (HD); slightly deleterious to slightly advantageous (SD-SA), or it can have a definitely positive, highly advantageous effect (HA). New mutants with fitness effects very close to zero will have evolutionary fates determined by random drift, they will be effectively neutral (red zone). In smaller populations, the range of mutants that can be classified as neutral is larger (grey zone), because weak selection becomes ineffective. The distribution of fitness effects presented here is over simplified, more detailed discussions can be found at [5] and [40].

lection to have a larger effect in organisms with relatively large population sizes. In populations where translational selection is ineffective, synonymous codons are expected to evolve neutrally. If we wish to understand why natural selection acts at the synonymous codon usage level in some organisms while ignoring others, we need to have a closer look at population genetic models of codon usage evolution. This is the topic of the next section.

2.2 Population genetics models of synonymous codon usage evolution

The following discussion is more mathematical in nature. The reader is now assumed to be familiar with the mathematical principles of population genetics. For those who wish to gain a quick but more formal and sound understanding of population genetics, the book by Hartl and Clark [53] is a good place to start.

The first population genetics approach to the problem of codon usage was made by Kimura [75], who developed a general model of a quantitative character evolving under stabilising selection, and applied this idea to the evolution of codon usage. In this model, the frequency of an optimal codon is the result of balancing selection driving the codon population to match that of the cognate tRNAs within the cell. Although this model might be mathematically right, it is not clear what biological argument justifies the assumption of balancing selection to maintain the frequency of a codon at an optimal level. Several workers used the idea of biased codon usage as evidence against the neutral theory [76]. It seems this might have troubled Kimura, who then tried (rather unsuccessfully) to accommodate the non randomness of codon usage into the neutral theory. As we shall see in the following section, the non randomness observed in codon usage bias for nearly every genome is perfectly compatible with the neutral theory. If selection is weak, and mutation against the optimal codon sufficiently strong, then the frequency of the optimal codon will

achieve an equilibrium that depends on the relative magnitude of these forces. A more formal treatment was later given by Li [89], who showed that strong codon bias could be achieved with small selection coefficients and large population sizes. However Li's work was focused more on computer simulations than on an analytical exploration of the problem. The first comprehensive treatment of the subject was given by Bulmer [13] who provided a complete population model to describe codon evolution. This is the model that will be discussed in detail in the following section. MacVean and Charlesworth [100] considerably extended Bulmer's work.

2.2.1 Bulmer's model of codon evolution

Let us consider a simple model where one amino acid a is encoded by only two synonymous codons, c_1 and c_2. The model is diploid, and c_1 is the optimal codon. The three genotypes have relative fitness $1+2s$ for the homozygote c_1c_1, $1+s$ for the heterozygote c_1c_2, and 1 for the homozygote c_2c_2, where s is the selection coefficient, *i.e.* the selective advantage, of c_1 over c_2. This is known as genic selection or semi-dominance. Let us have u as the mutation rate from c_1 to c_2 and v as the mutation rate in the reverse direction (mutations towards other amino acids is not allowed due to purifying selection). The frequency x of c_1 in a random mating population will vary from generation to generation due to the action of selection, mutation, and the random sampling of gametes due to finite population size. Thus x is a random variable and its probability density distribution [25] is given by

$$\frac{e^{Sx}x^{V-1}(1-x)^{U-1}}{\int_0^1 e^{Sy}y^{V-1}(1-y)^{U-1}dy}, \tag{2.1}$$

where $S=4N_es$, $V=4N_ev$, $U=4N_eu$ and N_e is the effective population size. When the mutation rates (u and v) are large compared to the effective population size, the density function above (equation 2.1) has a single mode clustered around the expected value of x and the codon location is expected to be polymorphic. When

the mutation rates are very small compared to the population size, the function has two sharp modes at zero and one, and the location is most likely to be fixed for either codon. The following discussion assumes the latter situation, and allows us to use fixation probabilities to calculate the value of x at equilibrium. The assumption of small mutation rates and little polymorphism seems reasonable for many proteins in natural populations [13]. A more complete model considering polymorphism is given elsewhere [100].

The probability of fixation of an advantageous mutant in a random mating population is approximately $\phi(S) = S/(2N(1-e^{-S}))$ when $|s|$ is small. Thus, every generation, a fraction $2Nux\phi(-S)$ of c_1 codons will mutate and become fixed towards c_2, and a fraction $2Nv(1-x)\phi(S)$ of c_2 codons will mutate and become fixed in the reverse direction towards c_1. At equilibrium $2Nux\phi(-S) - 2Nv(1-x)\phi(S) = 0$, from this the equilibrium value of x (x_{eq}) can be solved [89, 13] to give

$$x_{eq} = \frac{1}{1+\frac{u}{v}e^{-4N_e s}}. \tag{2.2}$$

This equation implies that the equilibrium frequency of the optimal codon depends on the relative mutation rate (u/v) rather than on the absolute mutation rates. Figure 2.5 shows the behaviour of x_{eq} for different values of u/v and $N_e s$. Note that if selection is sufficiently weak and mutation bias works against the optimal codon, then the most common codon might not be the optimal one.

Let us assume a gene g has n codon sites for amino acid a. The model assumes that each site evolves independently. It follows from the discussion above that at equilibrium, a fraction $n2Nux_{eq}\phi(-S)$ of the c_1 codon sites will be substituted with c_2, and a fraction $n2Nv(1-x_{eq})\phi(-S)$ of c_2 sites will be substituted with c_1. Thus the quantity

$$n\left(2Nux_{eq}\phi(-S) + 2Nv(1-x_{eq})\phi(S)\right)$$

represents the total number of substitutions per generation in gene g. The substitu-

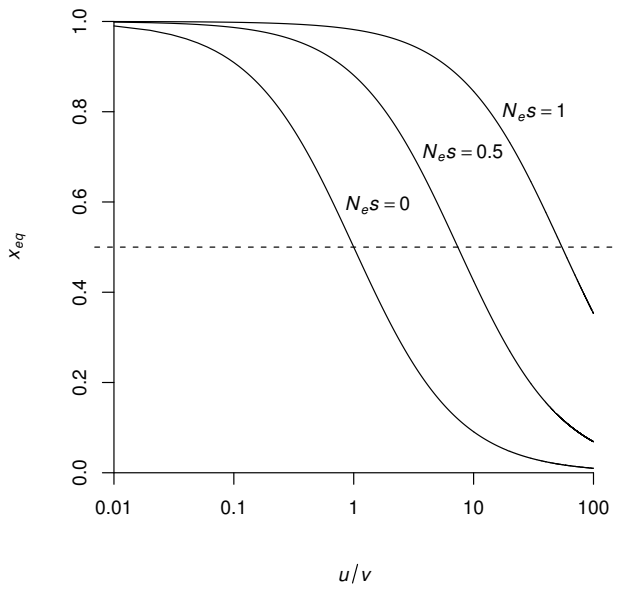

Figure 2.5: Frequency of an optimal codon at equilibrium

tion rate per site is then

$$\rho = 2Nux_{eq}\phi(-S) + 2Nv(1-x_{eq})\phi(S). \tag{2.3}$$

It can be shown that when $S = 0$, equation 2.3 is reduced to $\rho_0 = 2uv/(u+v)$. Thus $\rho_R = (\rho/\rho_0)$ is the relative substitution rate of a sequence under selection with respect to another one evolving neutrally [39]. The most important implication of equation 2.3 is that if mutational bias works against the optimal codon (*i.e.* $u > v$), then for small values of $N_e s$, the substitution rate is higher than for sequences evolving neutrally (*i.e.* $\rho_R > 1$, top figure 2.6). This also implies that for moderate values of x_{eq}, substitution rate and codon bias would seem uncorrelated (bottom figure 2.6). This has some important implications because there has been some controversy over whether substitution rates and codon bias are correlated or not [32]. It seems that for organisms where codon usage is moderate (such as *Drosophila* spp.), substitution rates and codon bias would seem uncorrelated, specially taking

into account that maximum likelihood methods that estimate substitution rates are subject to considerable error due to the high variances of the maximum likelihood estimates.

Bulmer's model assumes that every codon site in a sequence is essentially fixed, ignoring the problem of polymorphism at codon sites in a population. The derivation of equation 2.2 assumes that $4N_e u + 4N_e v \ll 1$. This seems to hold true for real populations, so most codon sites are fixed most of the time. However, small amounts of polymorphism do exist. MacVean and Charlesworth [100] developed a more comprehensive model of codon evolution that takes into account polymorphism at codon sites. Li [89] and later MacVean and Charlesworth [101] analysed the effects of linkage in codon evolution, and showed that this has dramatic effects on patterns of codon usage. The interested reader is strongly advised to consult these works for a deeper treatment of the subject.

2.2.2 Estimating S from actual data under Bulmer's model of codon evolution

Let us write $S = 4N_e s$ ($S = 2N_e s$ for an haploid organism) and $k = u/v$, then from equation 2.2 we have that

$$S = \ln\left(\frac{x_{eq}}{1 - x_{eq}} k\right), \tag{2.4}$$

as suggested by Bulmer [13] and later Sharp *et al.* [124]. Note that if $s = 0$ then equation 2.4 implies

$$k = \frac{1 - x_{eq}}{x_{eq}}.$$

Equation 2.4 provides an easy way to estimate S from actual data. Let us imagine a given organism where we wish to estimate S for a given amino acid encoded by two codons in a set of highly expressed genes. The mutation pattern is constant throughout the genome, codon frequencies are in equilibrium, and genes with low expression are under no selection for codon usage. Then an estimator $\hat{S}$ of S would

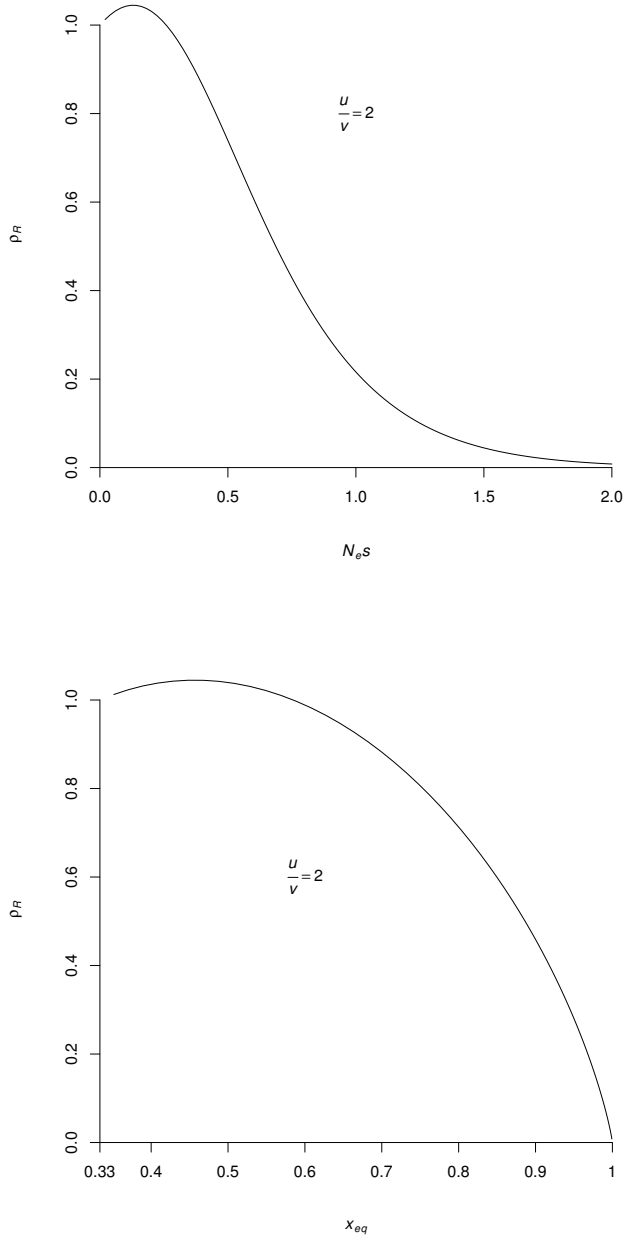

Figure 2.6: Substitution rate and selection on codon usage.
Top panel: Relative substitution rate (ρ_R) vs. $N_e s$. Note that ρ_R reaches its maximum value when $N_e s \approx 0.12$ and $u/v = 2$. For larger values of u/v the effect is more pronounced and the maximum moves towards the right. Bottom panel: ρ_R vs equilibrium frequency of the optimal codon (x_{eq}). Note that in the example shown, for values of x_{eq} between 0.3 and 0.6, substitution rate and codon bias seem uncorrelated.

be

$$\hat{S} = \ln\left(\frac{P_{hx}}{1-P_{hx}} \times \frac{1-P_{lx}}{P_{lx}}\right), \tag{2.5}$$

where P_{hx} is the observed frequency of the optimal codon in highly expressed genes and P_{lx} is the observed frequency in genes with low expression.

To exemplify how the technique above could be implemented, baker's yeast (*Saccharomyces cerevisiae*) genes were binned according to their expression level, and the frequencies of all the nine optimal codons belonging to the two synonymous family were computed. Optimal codons were taken as defined in the literature [10, 112, 68]. Expression data from microarray experiments [57] was used to partition the genes into 77 expression categories containing at least 6,000 codons. The $\hat{S}$ values for each of the nine optimal codons were estimated for each expression category. Genes expressed at 0.1 or less transcripts per cell were assumed to be under no selection and used to estimate k. Table 2.1 and figure 2.7 show the $\hat{S}$ values obtained. It can be seen that, as expected, average $\hat{S}$ values are positively correlated with expression level, with the largest $\hat{S}$ values showing an asymptotic relationship to expression level. An interesting observation is that there is considerable variation in $\hat{S}$ among genes (figure 2.7) and among amino acids (table 2.1). Eyre-Walker and Bulmer [39] reported similar trends for enterobacteria. In fact, how the set of highly expressed genes is constructed to estimate S can have substantial effects on the final estimated values. In our case, if a cut-off of ten transcript per cells is chosen to define highly expressed genes, then $\hat{S} = 0.75$, and indeed, selecting small random samples of genes with more than ten transcripts per cell would yield estimates of S anywhere between 0.75 and ~ 2.6. There has been some controversy over the wide discrepancies in $\hat{S}$ values for similar genomes in different studies (*e.g.* *Drosophila* spp. [97]), and these discrepancies could be partially explained by the use of different 'highly' expressed genes in different species (of the same genus) in these works. Considering that estimated effective population sizes for unicellular Eukaryotes are in the order of $10^7 - 10^8$ [92] and taking into account that yeast un-

dergoes haploid and diploid growth phases, the actual average selection coefficient (s) for codon usage in this organism would be between 6.5×10^{-9} and 1.3×10^{-7}.

This example serves to highlight the problem of estimating S in eukaryotic genomes. An important assumption is that mutation rates are constant throughout the genome. This assumption is not true since it is well known that there is large heterogeneity in mutation rates along genomes [91]. However, this assumption can be relaxed for baker's yeast, since its genome presents a narrow distribution of G+C content. Applying the approach described here to estimate S in larger Eukaryotes such as worm, fly or human, would be a greater challenge due to the large variation in G+C content seen in these genomes [14, 1, 85].

One final issue is the meaning of S. This is a confounded parameter that contains a numerical constant (2 or 4), the effective population size (N_e), and the actual selection coefficient (s). The value of the numerical constant depends on weather the organism is haploid or diploid, and in the case of diploidy, on the selection model being considered [25]. Nearly all studies seem to assume genic selection, if this assumption does not hold, then the exact form of S cannot be known unless a different model is explicitly specified. Furthermore, baker's yeast presents alternating haploid and diploid phases, which could arguably lead to oscillations between $2N_es$ and $4N_es$ depending on whether selection acts preferentially on the haploid or diploid phases. Thus the value of S estimated here includes a numerical constant somewhere between 2 and 4. However, it is important to note that S, as defined in equation 2.5, is simply the log odds ratio between the relative frequency of the optimal codon in highly vs lowly expressed genes. This is, nonetheless, a useful comparative measure of selected codon usage in disparate organisms such as bacteria, unicellular Prokaryotes or Metazoans.

Table 2.1: Estimated S values according to optimal codons in yeast.

	Codon[1]	P_{lx}	P_{hx}	u/v	$\hat{S}$[2]
Phe	TTT				
	TTC	0.38	0.86	1.66	2.34 (2.08, 2.63)
Tyr	TAT				
	TAC	0.41	0.92	1.41	2.78 (2.32, 3.43)
Cys	**TGT**	0.59	0.80	0.70	1.04 (0.50, 2.23)
	TGC				
His	CAT				
	CAC	0.34	0.79	1.91	1.95 (1.59, 2.26)
Gln	**CAA**	0.67	0.99	0.50	4.53 (3.69, 5.27)
	CAG				
Asn	AAT				
	AAC	0.37	0.91	1.67	2.79 (2.44, 3.16)
Lys	AAA				
	AAG	0.38	0.87	1.61	2.39 (2.19, 2.58)
Asp	GAT				
	GAC	0.34	0.60	1.95	1.08 (0.87, 1.27)
Glu	**GAA**	0.69	0.99	0.44	3.43 (3.03, 4.06)
	GAG				
Overall	-	-	-	-	2.57 (2.36, 2.78)

[1]Optimal codons are indicated with bold typeface.
[2]The two bins with higher expression values were pooled in order to obtain these estimates. Values in brackets are the bootstrap non-parametric 95% confidence intervals. See figure 2.7 for how the bootstrap values were obtained.

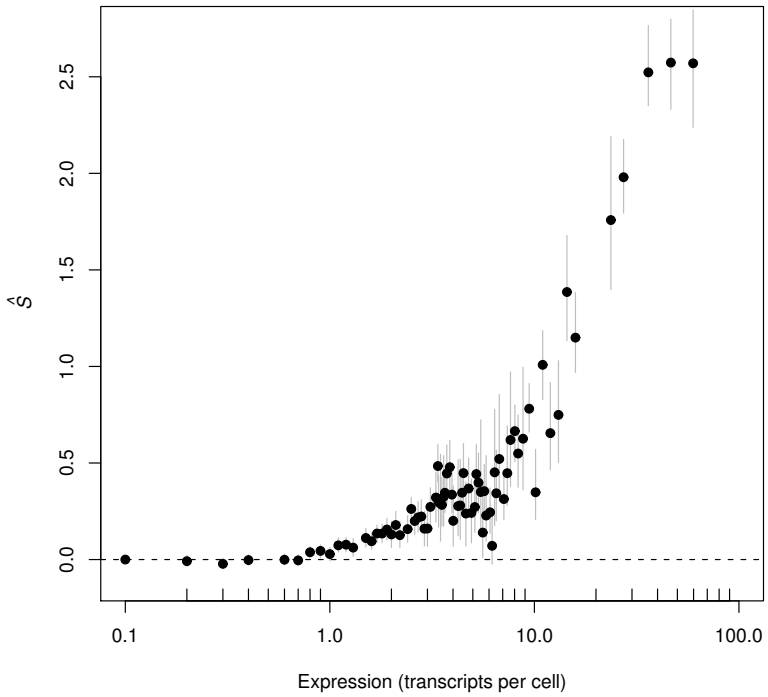

Figure 2.7: Expression levels vs average $\hat{S}$ values on codon usage for the baker's yeast.

The $\hat{S}$ values shown were averaged across all nine optimal codons and weighted by the total number of codons. Vertical bars are the non-parametric bootstrap 95% confidence intervals. Each bin contains at least 6,000 codons, and genes with the same expression value were binned together. Genes were sampled with replacement within each bin 1000 times, and the new samples were used to estimate the bootstrap values. Expression data obtained from [57]. This microarray set has been well studied in at least three separate publications on codon usage [21, 6, 7].

2.3 The nature of selection on codon usage

Translational optimisation means that optimal codons are those that match the most abundant cognate tRNAs within the cell. In cases where translational selection has been confirmed, it has consistently been noticed that synonymous codon usage bias in highly expressed genes resembles the configuration of the genomic set of cognate tRNAs [61, 112, 69, 33, 68]. This is the form of natural selection on codon usage that will be studied in more detail in the following chapters. However, this is not the only way natural selection can interfere with codon bias. For example, substitution rates at the beginning of enterobacterial genes are reduced [38], and this is attributed to selection for the maintenance of important regulatory elements that are located close to the translation initiation site. Similar constrains have also been reported in Eukaryotes [16, 17]. There are also reports of selection for secondary mRNA structure stability, or selection on codon usage associated to thermophilic lifestyle in bacteria [93]. These and other forms of selection that might act on codon usage are out of the scope of this work and will not be discussed here.

3 Measuring Codon Usage Bias

Several methods have been developed to measure codon usage bias, nearly all of them comprising some form of index that attempts to summarise in a single number the codon bias trends observed in a gene. Codon usage indexes have proved useful in several instances, and most of the literature on codon usage research is based on studies where one or more of these indexes have been applied to empirical data. The drawback has been that many of these studies have paid too much attention to the indexes themselves and little attention to the underlying population models that would explain the data.

Perhaps the simplest measure of codon usage bias is the frequency of optimal codons (F_{op}), first introduced by Ikemura [62]. Let op be the number of optimal codons and nop the number of non-optimal codons observed in gene g, then the frequency of optimal codons in g is defined as $F_{op} = op/(op + nop)$. This simple index has some nice properties, especially its direct relationship with the population genetic parameter S (chapter 2). In fact, for a sequence made up of only two synonymous codons, equation 2.5 can be rewritten as the log odds ratio of the frequency of its optimal codon

$$\hat{S} = \ln\left(\left(\frac{F_{op_{hx}}}{1 - F_{op_{hx}}}\right) \times \left(\frac{1 - F_{op_{lx}}}{F_{op_{lx}}}\right)\right),$$

where $F_{op_{hx}}$ is the frequency of the optimal codon in highly expressed genes, and $F_{op_{lx}}$ is that of genes with low expression. The above equation assumes equal mu-

tation rates in highly and lowly expressed genes, and that selection in genes with low expression is not operative. Expression levels correlate highly with F_{op}, as has been nicely shown by Akashi [6, 7] for the yeast genome. Despite its simplicity and easiness of computation, F_{op} seems to have lacked popularity.

Perhaps the two more popular codon usage indexes are the effective number of codons (Nc) [144] and the codon adaptation index (CAI) [128]. CAI forms the theoretical basis for the tRNA adaptation index developed in this chapter. Nc has some nice properties, especially that its expected value can be computed under the assumption of no selection, and the deviations from this assumption can be readily measured [144]. This latter fact forms the foundation for a technique developed in this chapter to detect translational selection in any completely sequenced genome. Both, CAI and Nc, have received the attention of several workers who have analysed the behaviour of these indexes in detail, and have proposed improvements and modifications [22, 64, 15, 44, 45, 46]. A myriad of codon usage indexes have been developed in more than twenty years of research, such as the scaled χ^2 [130], the z-value [69], maximum likelihood codon bias [136] and others. We will not detail any of these indexes here. The interested reader is referred to the work by Comeron and Aguadé [22] for a discussion of several of these indexes and their statistical performance.

3.1 The codon adaptation index

The codon adaptation index (CAI) was developed by Sharp and Li [128]. To calculate CAI for a group of genes in a particular organism, the set of optimal codons must first be determined. This is achieved by counting the number of occurrences of every codon in highly expressed genes (or ribosomal protein genes which are assumed to be always highly expressed, see [71]), and picking out the codons most frequently used for each amino acid. After constructing the set of optimal codons,

the *relative adaptiveness value* w_{ij} of codon i coding for amino acid j is defined as

$$w_{ij} = x_{ij} / \max_i(x_{ij}),$$

where x_{ij} is the absolute frequency of the i-th codon coding for the j-th amino acid in the set of highly expressed genes. Then, for any given gene g, its *codon adaptation index* CAI_g is defined as

$$\mathrm{CAI}_g = \left(\prod_{k=1}^{l_g} w_{i_k j_k} \right)^{1/l_g}, \tag{3.1}$$

where l_g is the length of the gene in codons, and $w_{i_k j_k}$ is the w-value for the i-th codon coding for the j-th amino acid defined by the k-th triplet in gene g. In order to overcome accuracy issues in computer calculations, equation 3.1 can be computed as

$$\mathrm{CAI}_g = \exp\left(\frac{1}{l_g} \sum_{k=1}^{l_g} \ln(w_{i_k j_k}) \right).$$

CAI is thus the geometric mean of the adaptiveness values of the codons present in gene g.

3.2 The effective number of codons

The effective number of codons (Nc) was developed by Wright [144], to measure the departure of a gene from equal usage of synonymous codons. It reaches is maximal value (61) when all codons are used equally and its minimal value (20) when only one codon is used per amino acid. Since the effect of selection is a reduction of the diversity of codon usage in a sequence, Nc provides a way of testing this effect. This index has been quite popular in codon usage research, particularly because it presents a well defined relationship to the silent GC content of a gene, so the effects of selection and mutational biases can be readily studied.

3.2.1 Definition of Nc

Wright used the analogy of multiple codons with multiple alleles [77], to calculate an homozygosity (F) value for the codons in a given amino acid. Let us assume an amino acid a that can be encoded by k different codons, and a gene g in which a appears n times, such that $n = n_1 + \ldots + n_k$, where $n_1, \ldots, n_k$ are the frequencies of the respective codons that encode a in g. The homozygosity F_a for amino acid a in gene g is defined as

$$F_a = \frac{n\sum_{i=1}^{k}(n_i/n)^2 - 1}{n-1}. \tag{3.2}$$

The effective number of codons for amino acid a, is then defined as

$$\mathrm{Nc}_a = \frac{1}{F_a}.$$

The value of Nc_a gives an idea of the codon usage diversity for amino acid a in gene g. For an amino acid encoded by four codons, Nc_a will range from 1, if only one codon is used to code for a, up to 4, if all codons are used equally. As seen in chapter 1 (table 1.1), amino acids can be classified according to their synonymous family (*i.e.* the number of synonymous codons that encode each particular amino acid). A homozygosity value ($\overline{F_i}$) for a given synonymous family i, can be computed as the average of all amino acids present in g and belonging to i. The overall Nc value of gene g is defined as

$$\mathrm{Nc}_g = 2 + \frac{9}{\overline{F}_2} + \frac{1}{\overline{F}_3} + \frac{5}{\overline{F}_4} + \frac{3}{\overline{F}_6}, \tag{3.3}$$

where $\overline{F}_i$ is the average homozygosity for those amino acids encoded by i synonymous codons. Note that the contribution of amino acids encoded by only one codon (*i.e.* Met and Trp) is set to two. The numerators in equation 3.3 reflect the number of amino acids in each synonymous family. There are nine amino acids encoded by two codons, one encoded by three codons and so on.

3.2.2 Nc and the silent GC content of a gene

It is useful to think of Nc_g as a random variable that changes over time as the codons in gene g change randomly due to mutation. If the mutation-drift process that generates the codons is stationary, then we expect Nc_g to oscillate around certain expected value $\mathrm{E}(\mathrm{Nc}_g)$ so that $\mathrm{Nc}_g = \mathrm{E}(\mathrm{Nc}_g) + \varepsilon_g$, where ε_g is the random deviation from the expected value due to the stochastic nature of Nc_g. Let us examine the behaviour of Nc for Phenylalanine ($\mathrm{Nc}_{\mathrm{Phe}}$) in a certain gene g. This amino acid is encoded by two codons only, UUC and UUU. The silent GC content of the genomic region where g is located is x, and the nucleotide frequencies in this region are in mutation-drift equilibrium. There is no translational selection acting on g, so the expected frequencies of UUC (z_1) and UUU (z_2) are simply x and $1-x$ respectively. Thus the expected homozygosity for the phenylalanine codons in g is

$$\mathrm{E}(F_{\mathrm{Phe}}) = \sum_{i=1}^{2} z_i^2 = x^2 + (1-x)^2,$$

and the expected Nc value for Phe is, for large n,

$$\mathrm{E}(\mathrm{Nc}_{\mathrm{Phe}}) = \frac{1}{x^2 + (1-x)^2}.$$

It is easy to verify that if $x = 0.5$ then $\mathrm{E}(\mathrm{Nc}_{\mathrm{Phe}}) = 2$ and if $x = 0$ or $x = 1$ then $\mathrm{E}(\mathrm{Nc}_{\mathrm{Phe}}) = 1$. A similar reasoning can be used to obtain $\mathrm{E}(\mathrm{Nc}_g)$ for a whole gene g, with GC3s content x_g, under the hypothesis of no selection

$$\mathrm{E}(\mathrm{Nc}_g) = f_{\mathrm{Nc}}(x_g), \tag{3.4}$$

where

$$f_{\mathrm{Nc}}(x_g) \;=\; 2 + \frac{19}{x_g^2 + (1-x_g)^2} + \frac{(2-x_g)^2}{2(1-x_g)^2 + x_g^2}$$

$$+\frac{27}{6\frac{3(x-x^2)^2+2x^2+(1-x)^4}{(1+x)^2}+x_g^2+(1-x_g)^2}. \tag{3.5}$$

A complete derivation of this expression is presented in the appendix on page 133.

Another solution to the expected value of Nc has been proposed by Wright [144] as

$$\mathrm{E}(\mathrm{Nc}_g) = 2 + x_g + \frac{29}{x_g^2+(1-x_g)^2}. \tag{3.6}$$

This formula gives a reasonable approximation, although Wright did not provide a derivation for it. A different definition of N_c that should be insensitive to the background nucleotide composition has been proposed by Novembre [105].

Equation 3.4 can be used to construct the so called Nc-plots [144]. An Nc-plot for a sample of genes is constructed plotting the Nc values obtained for each gene against their particular GC3s content, and adding the curve produced by equation 3.4 into the graph. Nc-plots are very useful as a descriptive analysis of codon usage in any given genome because the deviation of the genes from their expected Nc value can be readily observed. Figure 3.1 shows an Nc-plot depicting some yeast genes and some simulated genes. The simulated genes were generated under the hypothesis of no selection, and it can be seen that they follow the theoretical expectation proposed herein (equation 3.5). The yeast genes are split into two groups, a large cluster that falls largely over the theoretical expectation, and a smaller set of genes that present Nc values that are much smaller than expected by chance. These genes are highly expressed and use a small set of optimal codons that match the most abundant tRNAs within the yeast cell.

3.3 The tRNA adaptation index

An alternative way to study codon usage bias is to analyse the adaptation of a particular gene to the tRNA pool of its genome. If the codon usage pattern of a particular

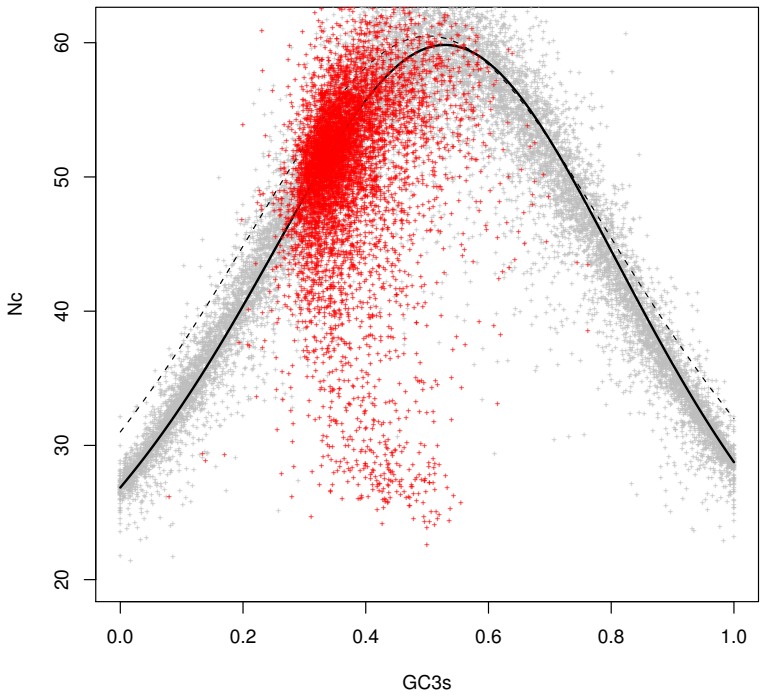

Figure 3.1: Nc-plot for yeast and simulated *E. coli* K12 genes.
Dashed line: Wright's function (equation 3.6); bold line: the function proposed herein (equation 3.4) . Red points: actual yeast genes; grey points: simulated *E. coli* K12 genes. Each open reading frame present in the *E. coli* K12 genome was simulated three times, and the simulated genes were generated according to the following rules: (i) the amino acid composition should remain intact, (ii) the codon that codes for any given amino acid was chosen randomly according to a set silent GC content for its gene, and (iii) the silent GC content of any given gene was chosen randomly from a uniform distribution.

gene matches the most abundant tRNAs, this can be interpreted as strong evidence of the action of translational selection. The first person who seems to have used this approach was Ikemura [61] in his analysis of tRNA expression levels and codon usage in *E. coli*. Ikemura simply measured the usage frequency of tRNA species in particular genes (for a given tRNA species which recognises one or more codons, its usage frequency is defined as the overall frequencies of the codons themselves) and plotted these values against the expression level for each tRNA species as measured by two dimensional gel electrophoresis. With this simple approach, Ikemura convincingly showed that selection was operative at the synonymous codon level in order to optimise the *E. coli*'s translational system [60]. A more refined approach is to modify Sharp and Li's CAI [128], so that the relative adaptiveness of each codon is computed not from the frequency of codons in highly expressed genes, but from the expression levels of the corresponding tRNAs. This tRNA adaptation index (tAI), was first implemented by dos Reis *et al.* [30] using genomic tRNA gene copy number as a surrogate of actual tRNA expression levels. The main advantage of tAI over CAI is that it can be computed for any arbitrary genome where the genomic tRNA pool is known, and where the determination of the optimal codons from highly expressed genes is unclear.

3.3.1 Definition of tAI

In order to develop tAI, we take advantage of the fact that tRNA gene copy number across some genomes correlates highly and positively with tRNA abundance within the cell and with codon preferences in such genomes [61, 69, 112, 33]. Since tRNA abundance might be thought of as the driving force for translational selection, it is reasonable to speculate that measuring the tRNA usage of a gene might provide an indirect way for detecting translational selection according to how well the gene in question is adapted to the tRNA gene pool. In order to calculate this index, the absolute adaptiveness value W_i for each codon i is defined as

$$W_i = \sum_{j=1}^{n_i} (1 - s_{ij}) \text{tGCN}_{ij}, \tag{3.7}$$

where n_i is the number of tRNA isoacceptors that recognise the i-th codon, tGCN_{ij} is the gene copy number of the j-th tRNA that recognises the i-th codon, and s_{ij} is a selective constraint on the efficiency of the codon-anticodon coupling. To build a table of W_i values, it is best to sort the codons as shown in figure 3.2 and to analyse the way in which each tRNA recognises its particular codons. It can be seen that the 64 codons that comprise the genetic code can be clustered into groups of 4 elements, which reflect the natural way in which tRNAs recognise them. Based on figure 3.2, a simple set of formulae for calculating all W_i values can easily be drawn taking into account Crick's wobble rules [24] for codon-anticodon pairing (table 3.1). From values W_i the *relative adaptiveness value* w_i of a codon is obtained as

$$w_i = \begin{cases} W_i / W_{\max} & \text{if } W_i \neq 0 \\ w_{\text{mean}} & \text{else,} \end{cases} \tag{3.8}$$

where $W_{\max}$ is the maximum W_i value and w_{mean} is the geometric mean of all w_i with $W_i \neq 0$. The *tRNA adaptation index* tAI_g of a gene g is defined as the geometric mean of the relative adaptiveness values of its codons:

$$\text{tAI}_g = \left(\prod_{k=1}^{l_g} w_{i_k} \right)^{1/l_g}, \tag{3.9}$$

where i_k is the codon defined by the k-th triplet in gene g and l_g is the length of the gene in codons (except the stop codon). Consequently, tAI_g estimates the amount of adaptation of a gene g to its genomic tRNA pool.

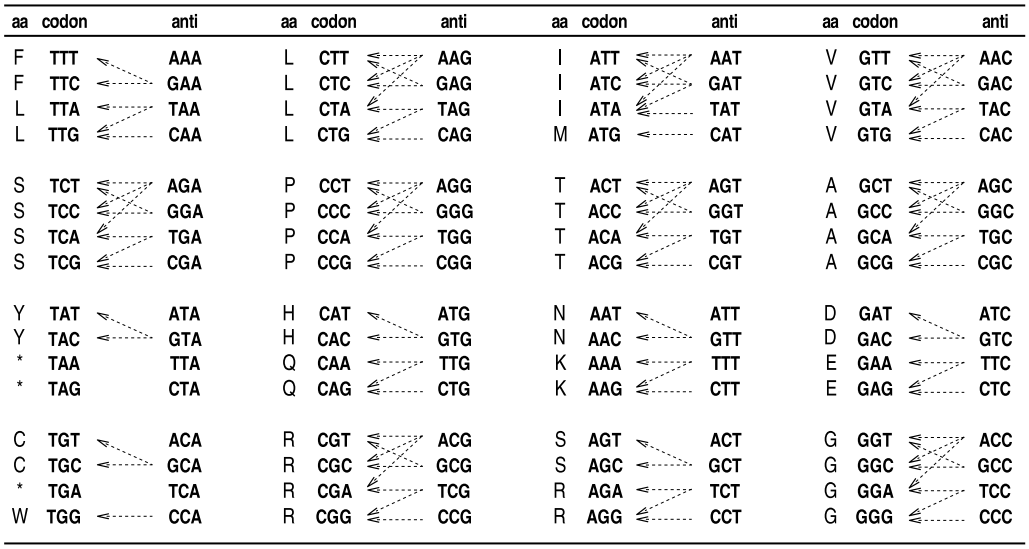

Figure 3.2: General codon-anticodon recognition rules for tRNA genes.
This table simply summarises all the theoretically possible interactions between the coding codons and the extant tRNA sequences in the organisms analysed in this work. The interested reader is advised to consult the literature [139, 155] for a detailed description of codon-anticodon pairings.

Table 3.1: Formulae for calculating W's according to Crick's wobble rules.

n	Anti-codon	Codon	W
i	INN	NNU	$(1-s_{I:U})tGCN_i + (1-s_{G:U})tGCN_{i+1}$
i+1	GNN	NNC	$(1-s_{G:C})tGCN_{i+1} + (1-s_{I:C})tGCN_i$
i+2	UNN	NNA	$(1-s_{U:A})tGCN_{i+2} + (1-s_{I:A})tGCN_i$
i+3	CNN	NNG	$(1-s_{C:G})tGCN_{i+3} + (1-s_{U:G})tGCN_{i+2}$

I: inosine. The interested reader should consult the literature [139, 155] for a detailed description of nucleoside modifications and codon-anti-codon pairings. Crick's wobble rules are in [24].

3.3.2 Choosing appropriate *s*-values for calculating tAI

One of the most challenging issues when computing tAI is the selection of a meaningful set of s_{ij}-values (equation 3.7). Since tRNA usage should be maximal for highly expressed genes, it would be natural to find the set of s_{ij}-values that maximise the correlation between expression levels and tAI values for any given organism. Microarray data from yeast (figure 2.7 on chapter 2), was used to optimise these values. A set of highly expressed genes was selected using the criteria chosen previously [21], and tAI was calculated for every one of them, assuming initial s_{ij}-values as shown in table 3.2. The correlation between the obtained tAI values for each gene and its corresponding expression level was then calculated iteratively using an implementation of the Nelder and Mead algorithm (R package) until the optimal set of s_{ij}-values that maximised this correlation (here, $R_{final} = 0.71$) was obtained (table 3.2).

A similar method was used with *E. coli* microarray data [11] to obtain *s*-values for prokaryotic organisms (table 3.2), with special attention to the recognition of AUA by LAT ($s_{L:A}$) in these genomes (table 3.2). It is important to note that when this analysis was originally performed, high quality microarray data for *E. coli* (or other prokaryotes) was not available [30]. For several microarray data sets analysed, dos Reis [30] found a very weak correlation between expression level and codon bias. Thus the optimisation procedure for *E. coli* is unstable since substantial variations in the s_{ij}-values have relatively little effect on the correlation between tAI and expression. For this reason, using these s_{ij}-values (with the exception of $s_{L:A}$) when computing tAI for prokaryotic organisms is not recommended. The yeast s_{ij}-values should be used instead. Whether there should be any differences in the recognition efficiency (i.e. the s_{ij}-values) of particular tRNA anticodon-codon pairs in Eukaryotes and Prokaryotes is not clear. This is an interesting topic that should be further investigated in the future. Researchers interested in computing accurate tAI values for particular genomes are advised to obtain reliable expression data for

Table 3.2: Optimised s-values

s	fixed	s	initial	final (yeast)	final (*E. coli*)
$s_{I:U}$	0.0	$s_{G:U}$	0.50	0.41	3×10^{-7}
$s_{G:C}$	0.0	$s_{I:C}$	0.50	0.28	0.78
$s_{U:A}$	0.0	$s_{I:A}$	0.25	0.9999	4×10^{-6}
$s_{C:G}$	0.0	$s_{U:G}$	0.50	0.68	0.86
-	-	$s_{L:A}$	0.50	-	0.89

L: Lysidine.

the organism in question and perform the optimisation of s_{ij}-values as described above.

3.3.3 Relationship of tAI to Nc

As we have seen above, tAI measures the adaptation of the codons in a gene to the genomic tRNA pool. If we observe a gene with a high value of tAI, does this mean its codon usage has been shaped by natural selection to match the tRNA set? If we measure tAI for a large set of genes in a given genome, how do they relate to the genomic tRNA pool? Is there any evidence of the action of translational selection? I believe the key to these questions lies in the relationship of tAI with Nc. As we have seen, Nc has a theoretically expected relationship to the silent GC content of a gene. When a set of genes present values of Nc substantially lower than those expected from their GC3s content, this is suggestive of the action of natural selection. If we could quantify this departure, and prove that it correlates with the tRNA usage of the genes in question, then a rough measure of the strength of translational selection in this set of genes could be obtained. A theoretical basis for such test is laid out below.

Let us imagine a lowly expressed gene g_{lx} , from a particular organism, is in mutation-drift equilibrium and selection on codon usage is not operative. This gene will have an expected Nc value given by equation 3.4. Now let us imagine a highly expressed gene g_{hx} in the same genome, that is under the strong action of transla-

tional selection. It is assumed that both genes are under the same type of mutational biases. Both genes could be placed in an Nc-plot (figure 3.3). We would notice that the Nc value of g_{hx} ($\mathrm{Nc}_{g_{\mathrm{hx}}}$) would be substantially different from that of its lowly expressed counterpart ($\mathrm{Nc}_{g_{\mathrm{lx}}}$). Because selection reduces the overall codon diversity of a sequence, we would expect $\mathrm{Nc}_{g_{\mathrm{hx}}} < \mathrm{Nc}_{g_{\mathrm{lx}}}$ to be true. Thus, we define the effect of selection (ψ) on the Nc value of g_{hx} as $\psi_{g_{\mathrm{hx}}} = \mathrm{Nc}_{g_{\mathrm{lx}}} - \mathrm{Nc}_{g_{\mathrm{hx}}}$. Selection might have also altered the GC content of g_{hx} compared to its lowly expressed counterpart. For example, if g_{hx} is in an AT rich genome, and optimal codons end in C or G, then the silent GC content of g_{hx} would have been increased due to the action of selection. If x_{lx} is the GC3s content of g_{lx} and $x_{g_{\mathrm{hx}}}$ is the GC3s content of g_{hx}, then, the effect of selection (ω) on the GC3s content of g_{hx} is defined as $\omega_{g_{\mathrm{hx}}} = x_{g_{\mathrm{hx}}} - x_{g_{\mathrm{lx}}}$.

Considering the example above, it is easy to see that the expected Nc value of g (equation 3.4) can be extended to account for the effect of selection on (ψ_g) and GC3s content (ω_g), compared to the expected state under no selection,

$$\mathrm{E}(\mathrm{Nc}_g) = f_{\mathrm{Nc}}(x_g - \omega_g) - \psi_g, \tag{3.10}$$

and the *actual* value of Nc for g would be

$$\mathrm{Nc}_g = f_{\mathrm{Nc}}(x_g - \omega_g) - \psi_g + \varepsilon_g, \tag{3.11}$$

where ε_g is the random deviation from the expected value.

In the following discussion, it will be assumed that selection has little or no effect on GC3s content, this means $\omega_g \approx 0$ so that $f_{\mathrm{Nc}}(x_g - \omega_g) \approx f_{\mathrm{Nc}}(x_g)$. As we shall see, this assumption seems to be reasonable for several genomes. The implications of values of ω_g substantially larger than zero will be discussed later when the method being developed here is applied to real data. The assumption above means that equation 3.11 can be reduced to

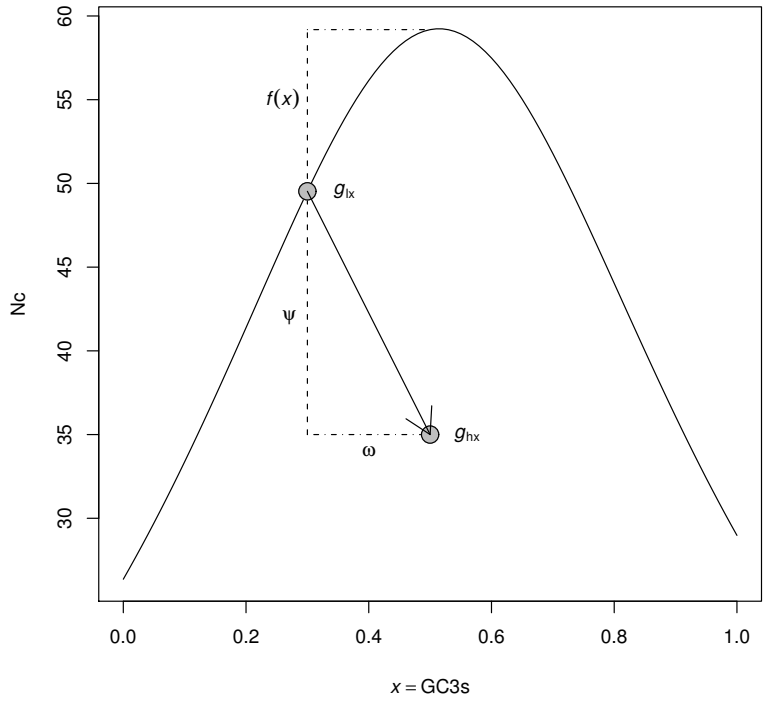

Figure 3.3: The action of translational selection on a gene in an Nc-plot.
A lowly expressed gene g_{lx} is in mutation-drift equilibrium with no selection. Another highly expressed gene g_{hx}, in the same genome, has been modified by the action of natural selection and its codon usage is in a highly optimised state. The arrow represents the action of selection on g_{hx}, which can be decomposed into two components: ω which is its effect on the value of GC3s, and ψ which is its effect on the value of Nc. The term $f(x)$ represents the reduction in the expected value of Nc due to GC3s content in the absence of selection.

$$\mathrm{Nc}_g = f_{\mathrm{Nc}}(x_g) - \psi_g + \varepsilon_g. \tag{3.12}$$

The random element ε_g simply represents sources of variation on Nc that cannot be accounted for by selection or silent GC content alone, and that cannot be controlled in this study. Since in practical terms we can only calculate Nc_g and $f_{\mathrm{Nc}}(x_g)$ for every gene, terms ψ_g and ε_g in equation 3.12 are confounded, and cannot be estimated independently. Consequently, the amount of selection acting on the codon usage of gene g cannot be estimated directly, but the confounded factor, $\varphi_g = \psi_g - \varepsilon_g$ can be estimated as

$$\varphi_g = f_{\mathrm{Nc}}(x_g) - \mathrm{Nc}_g. \tag{3.13}$$

We can use tAI_g and φ_g to estimate the degree of co-adaptation between the codon usage of a set of genes and the genomic tRNA pool. If a representative sample of genes from a given genome G is obtained, then the vectors $\mathrm{tAI}_G = (\mathrm{tAI}_g)$ and $\Phi_G = (\varphi_g)$ can be calculated. The correlation S_t between tAI_G and Φ_G measures this co-adaptation. In fact, the squared correlation coefficient S_t^2 is the proportion of the variance in codon bias (as measured by Nc) explained by tRNA adaptation that cannot be explained by GC content variation (x_g) or other factors (ε_g) alone. The correlation S_t is a convenient indicator of the amount of selection due to tRNA adaptation since it is a single number between -1 and 1. It can be seen that the stronger the action of selection, the higher the correlation coefficient. If this test is applied to a representative set of genes in any given organism, a measure of the intensity of translational selection that has acted upon the evolution of its genome will be obtained. It is important to distinguish S_t here from $S = 4N_e s$ (chapter 2). The value of S_t simply summarises how strong selection has been on a whole genome in optimising its codon usage, and S_t cannot be thought of as an estimator of the popu-

lation genetic parameter S[1]. It is reasonable to expect that S_t is somehow related to S in an actual genome, however an analytical treatment of the subject is out of the scope of this work. An empirical relationship between estimates of S_t and S from actual genomes is discussed in chapter 4.

An example of how the S_t test is implemented is shown in figure 3.4. The bacterium *E. coli* shows Nc values that deviate strongly from their theoretically expected values under no selection. These deviations can be explained by co-adaptation of the codon bias to the genomic tRNA set, hence, this genome presents a substantially large S_t value. The opposite case is exemplified in the genome of *H. sapiens*, where Nc values closely follow the theoretical expectation. The small deviations actually observed are due to different content of G vs. C and A vs. T. The statistical significance of the S_t values observed in these genomes can be assessed by a permutation test. The method consists in permuting the assignment of w_i values to their respective codons. The permuted set is then used to calculate tAI and S_t, and the process is repeated iteratively until a sufficiently large sample of S_t values is generated to estimate its probability distribution under the assumption of no selection. Thus p-values for the significance of the naturally observed S_t values can be obtained from this re-sampling distribution. *E. coli* shows a highly significant S_t value ($p \approx 0.003$) while *H. sapiens* shows no evidence of codon-tRNA coadaptation ($p \approx 0.114$).

Recently, Man *et al.* [95, 94] carried out an extensive analysis of the performance of tAI on ten yeast genomes. They found that this index correlates with protein expression levels, even after correcting for mRNA levels, *i.e.* tAI can explain part of the variation of protein expression for genes having very similar mRNA levels.

[1] When originally published, the S_t parameter was simply represented as S for 'selection' without any population genetics justification [28]. The year after the original work was published, Sharp *et al.* [124] presented their formal population genetics measure, S, for the strength of selected codon usage. For discussion purposes in their work, and to distinguish my index from their population genetics parameter, they renamed my S parameter as S_t (which I would like to think stands for 'selection from tRNA'). In any case Sharp *et al.* [124] nomenclature is adopted here despite the confusion it may cause.

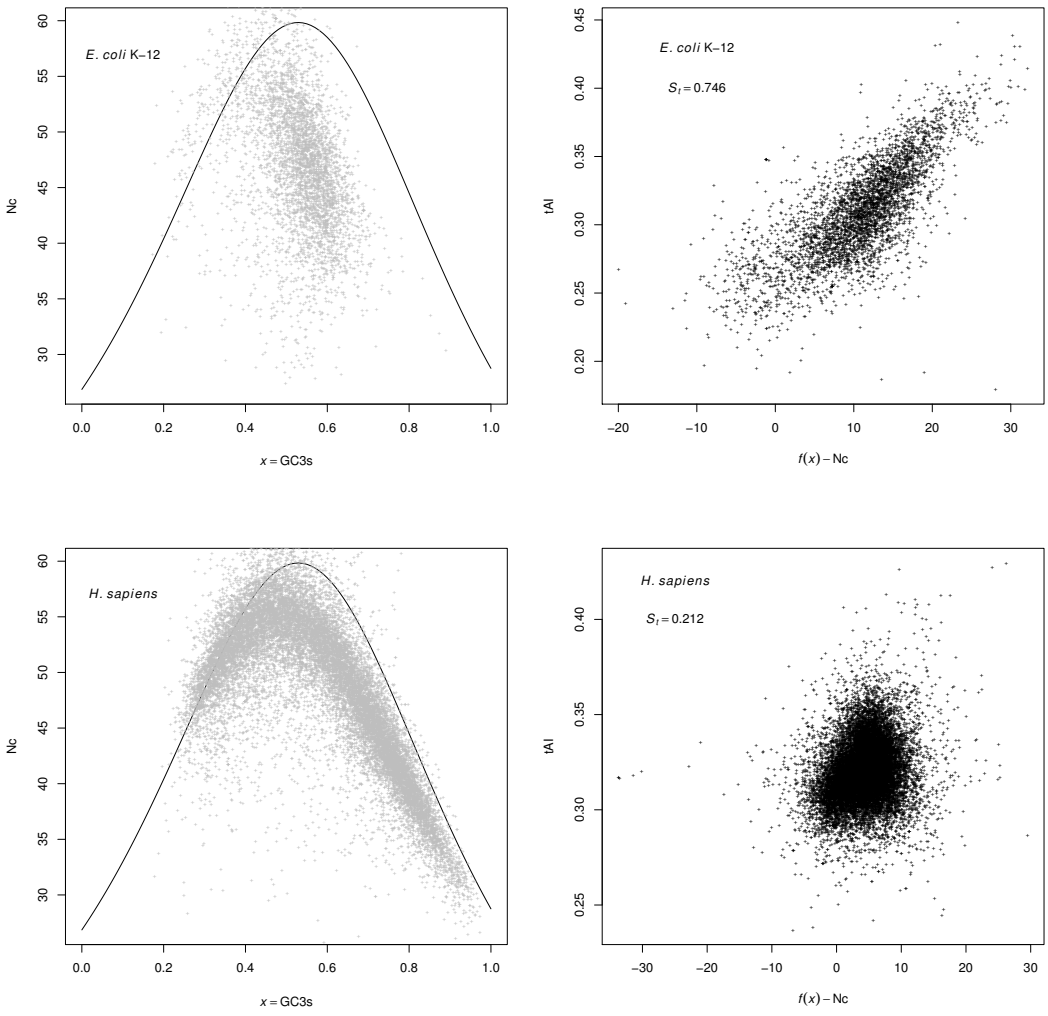

Figure 3.4: S_t test for *Escherichia coli* K-12 and *Homo sapiens.*
Nc-plots (left panel) and S_t plots (right panel) are presented. S_t values were calculated only for genes with more than 100 codons. Human genes do not follow the theoretical expectation for N_c perhaps due to the strong variation in CpG dinucleotide content [85].

They also provide a worthwhile discussion of the advantages of tAI and a comparison with CAI. These workers also used an analogous way to the S_t test to determine whether selection has been operative in the ten genomes analysed. They simply calculated the correlation between tAI and Nc (without correcting for GC3s content), and showed that translational selection is operative in at least nine of the ten yeast genomes analysed.

Now that we have a rough method to detect whether translational selection has been operative in any fully sequenced genome, we shall apply it to a large data set of eukaryotic and prokaryotic organisms. This is the topic of the next chapter.

4 Trends of codon-tRNA coadaptation in eukaryotic and prokaryotic genomes

As discussed in chapter 1, one of the challenging issues in codon usage research is to understand why natural selection optimises codon usage in certain genomes while ignoring others. Now that we have developed a method to test whether the genomic tRNA set of an organism explains part of the variation in the codon usage of its genome, it would be natural to apply this test to a broad sample of Prokaryotes and Eukaryotes in order to tackle the question above. This chapter explores the degree of codon usage-tRNA coadaptation in several prokaryotic and eukaryotic genomes from the point of view of the S_t test.

4.1 Estimates of S_t in several prokaryotic and eukaryotic genomes

One hundred and thirty-two genomes, from Archaea to Eukaryota, were analysed for signs of natural selection on codon usage. Table 4.1 (p. 81) shows the genomes considered, their estimated S_t values, and other genomic variables of interest. A phylogenetic tree showing the main relationships among the groups of organisms analysed is depicted in figure 4.1. The organisms analysed ranged in genome size

from 0.58 Mb (*Mycoplasma genitalium*) up to around 3,000 Mb (*Homo sapiens*), with total tRNA gene copy numbers ranging from 29 (*Mycoplasma pulmonis*) to 620 (*Arabidopsis thaliana*). The presence or absence of translational selection seems to be independent of the kingdom being considered, both Eukaryotes and Prokaryotes presented organisms whose codon usage is largely explained by selection or mainly by mutational processes. S_t values ranged from -0.30 (*Halobacterium* sp.) up to 0.86 (*Cryptococcus neoformans*). In total, 43 genomes have values of S_t statistically different from zero ($P < 0.05$).

An interesting trend that can be inferred from observing figure 4.2 is that organisms with intermediate genome sizes and large tRNA numbers (> 50) tend to show larger S_t values. These trends can also be seen when each superkingdom (Archaea, Bacteria and Eukaryota) is analysed separately. Figure 4.3 shows the relationship between S_t and tRNA gene number for each domain. In prokaryotic organisms (Archaea and Bacteria), S_t increases with increasing tRNA gene numbers. Rocha [120] and Sharp *et al*. [124] have suggested that in Bacteria, selective pressure for fast growth rates might relate to tRNA content, with fast growing bacteria presenting the largest number of tRNA genes. We discuss the relationship between growth rate and S_t in bacteria in section 4.4. The opposite trend, a reduction of S_t with tRNA number, is actually observed for eukaryotic organisms. The yeast genomes, which present moderate tRNA numbers (~200 gene copies) show the largest values of S_t (table 4.1) with S_t decaying with increasing redundancy in the tRNA set. Interestingly, the two parasitic Eukaryotes with the lowest number of tRNA genes (*Encephalitozoon cuniculi* and *Plasmodium falciparum*, table 4.1) show low S_t values, resembling those of bacteria with similar tRNA numbers (figure 4.5, top).

Interesting trends can also be noticed when S_t is compared to genome size for each superkingdom (figure 4.4). In Archaea, S_t increases with increasing genome size. For bacteria, the trend is less clear. The largest values of S_t are achieved in bacteria with moderate genome sizes (~5Mb) such as *Escherichia coli* (table 4.1).

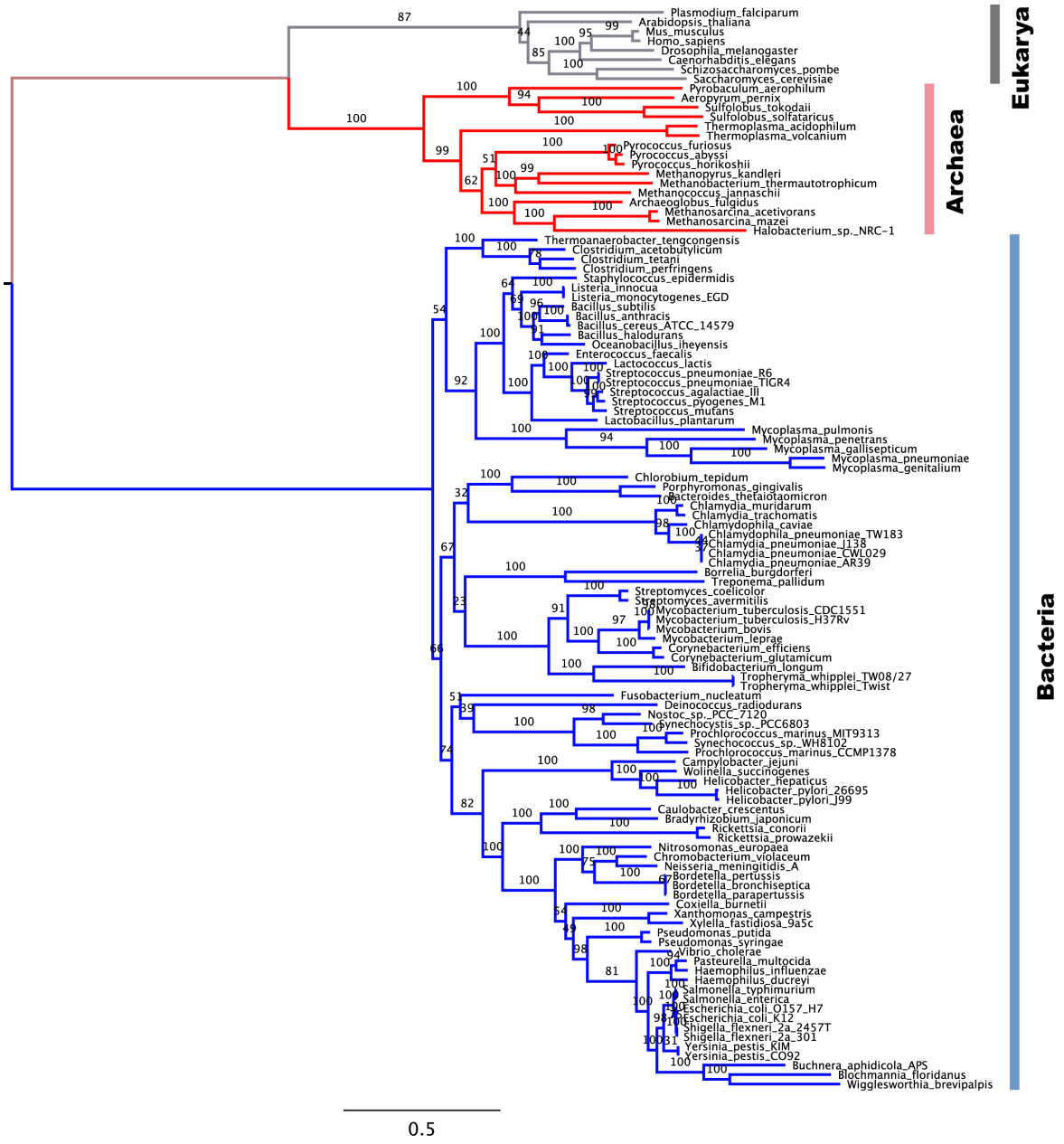

Figure 4.1: Phylogenetic tree depicting the relationships among the main groups of organisms analysed.

The tree topology was estimated by Ciccarelli *et al*. [19] for 191 genomes. They used a sophisticated approach to identify a set of 31 orthologous genes conserved across the three domains of life. Their method allowed the estimation of branch lengths (in substitutions per site, scale bar) that are comparable across the different groups. Numbers on branches are the bootstrap support values. The tree of Ciccarelli *et al*. [19] was compared to the 132 organisms studied here, and the tree was pruned down to a set of 114 common organisms between both studies. This smaller tree was used to calculate phylogenetically independent contrasts (PIC) for the variables of interest [42, 110]. PIC were calculated with the Ape package for the R statistical environment [111].

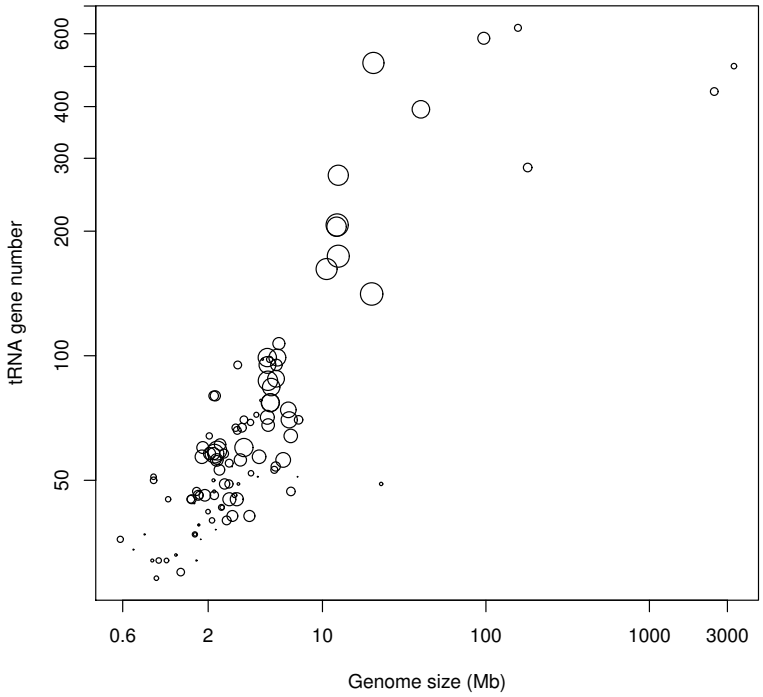

Figure 4.2: S_t, tRNA gene number and genome size for several eukaryotic and prokaryotic organisms.
Each plotting symbol diameter is proportional to the S_t value of the genome being represented. It can be seen that S_t values tend to be larger (bigger circles) for genomes of intermediate size (~ 2 - 100 Mb). The correlations among the PIC for tRNA number vs. log(genome size) for each superkingdom are $R_{\mathrm{Archaea}} = 0.90$, $R_{\mathrm{Bacteria}} = 0.69$ and $R_{\mathrm{Eukaryota}} = 0.85$.

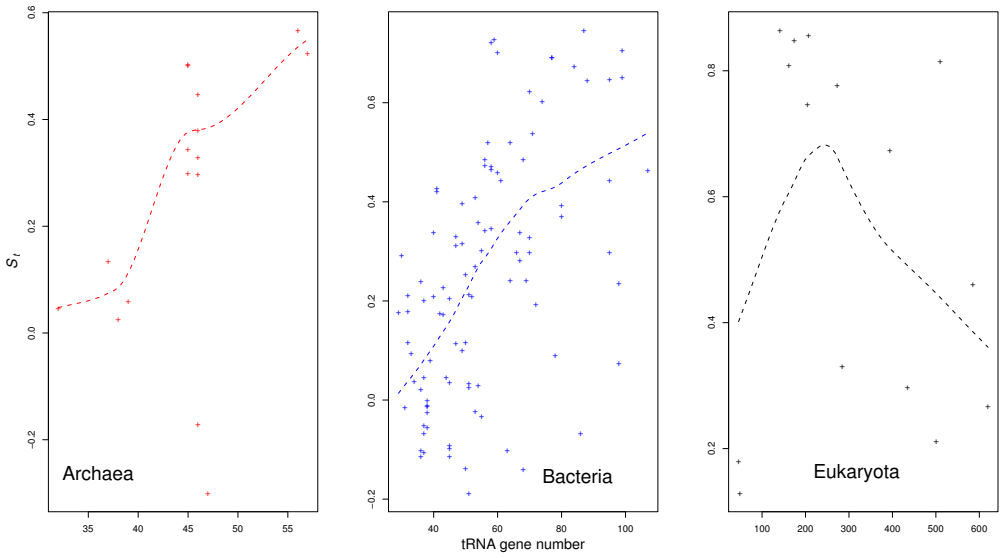

Figure 4.3: S_t vs. tRNA gene number for each superkingdom.
The regression lines are loess smooth regressors [138]. The regression is by a local polynomial, with the value of the function at point x being weighted according to the neighbours of x. The correlations between the PIC for S_t vs. tRNA number for each super kingdom are: $R_{\text{Archaea}} = 0.60$, $R_{\text{Bacteria}} = 0.084$ and $R_{\text{Eukaryota}} = -0.42$.

The largest bacterial genomes (~10Mb) show low S_t values. In Eukaryotes, a decrease in S_t values is observed with increasing genome size. Yeast genomes with the largest S_t values have moderate genome sizes (~10-20Mb). Interestingly, *Encephalitozoon cuniculi* which presents the smallest genome size (~2Mb) for the Eukaryotes analysed, presents S_t values resembling those of bacteria with similar genome size (figure 4.5, bottom).

When S_t is compared to tRNA gene number and genome size for all superkingdoms together (figure 4.5), a bell shaped relationship between S_t and these variables is apparent. Figures 4.3 and 4.4 are orthogonal projections of S_t on tRNA numbers and genome size. It might be useful to construct a regression surface of S_t on these two variables simultaneously as this might clarify the trends observed when the variables are considered separately. To achieve this, a nonparametric regression of S_t on tRNA numbers and genome size was performed. A Gaussian processes (GP) model was applied to the data since it provides a very flexible and powerful way to

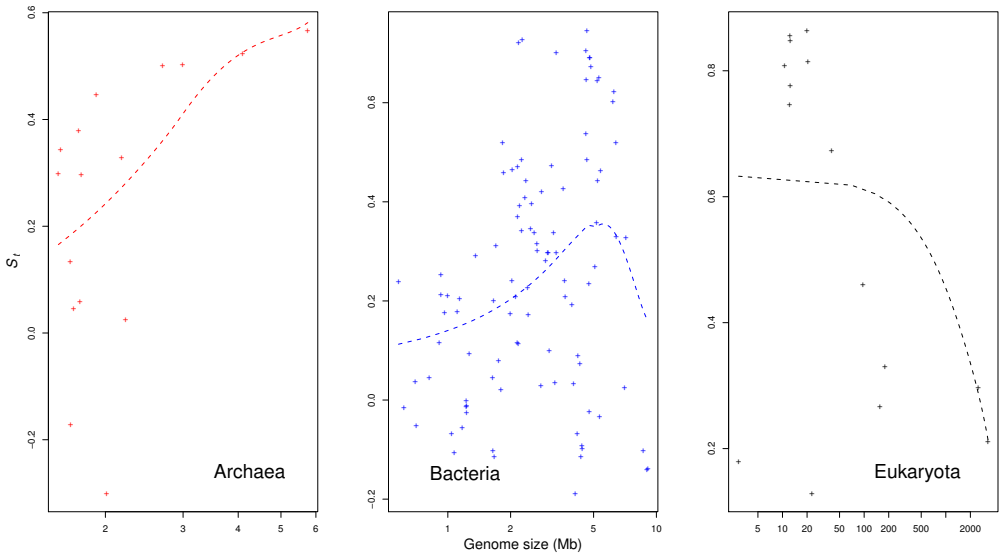

Figure 4.4: S_t vs. genome size for each superkingdom
The regression lines are loess smooth regressors [138]. The correlations between the PIC for S_t vs. log(genome size) for each superkingdom are: $R_{\text{Archaea}} = 0.79$, $R_{\text{Bacteria}} = 0.50$ and $R_{\text{Eukaryota}} = -0.055$.

analyse data with unusual properties such as correlated predictors (like genome size and tRNA gene number in this case). Gaussian processes are specified by a covariance structure that determines how response values at neighbouring points influence each other. A fully Bayesian treatment of the parameters describing the covariance structure is possible by a Markov Chain Monte Carlo (MCMC) algorithm. Random surfaces are generated according to how well they fit the observed data points, and a consensus regression surface is then obtained by averaging. Gaussian processes are nowadays considered a standard regression technique [138]. The details about how the covariance structure of the errors is constructed, and the associated likelihood function and MCMC analysis is described in the excellent introduction to Gaussian processes by Rasmussen and Williams [118].

The bell-shaped surface observed in figure 4.6 is the consensus obtained from the MCMC GP. A logarithmic transformation was used to scale tRNA gene numbers and genome sizes to appropriate values in the regression analysis. Also, a generalised additive model (GAM) [55] was fitted to the data as an independent re-

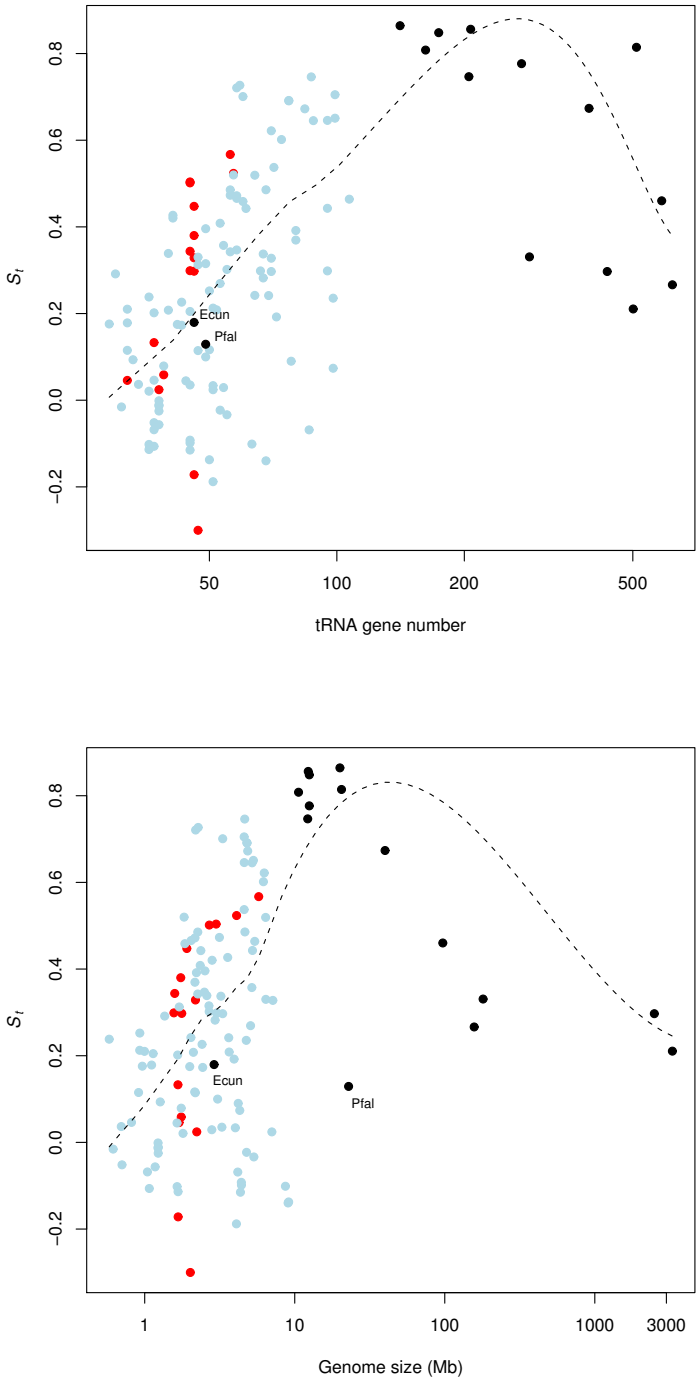

Figure 4.5: S_t vs. tRNA number and S_t vs. genome size for all organisms combined. Top: S_t vs. tRNA gene number for all three superkingdoms. Bottom: S_t vs. genome size for all three superkingdoms. The regression lines are loess smooth regressors [138], and were fitted for all three superkingdoms simultaneously. Red: Archaea; Blue: Bacteria; Black: Eukaryotes. Note the position of the two intracellular parasitic Eukaryotes (*Encephalitozoon cuniculi*: Ecun; and *Plasmodium falciparum*: Pfal).

gression model to confirm the quality of the analysis. The GAM surface indicates that tRNA gene number and genome size explain about 60% of the variation observed in S_t values. Both, the GAM and the GP surfaces are strikingly similar, but the Gaussian model was preferred because the Bayesian approach integrates appropriately over the uncertainty in all the model parameters. The fitted GP surface can be represented as a thermal image (figure 4.7) that shows the hot regions of codon usage-tRNA coadaptation. This landscape shows a conspicuous hot spot where the activity of selection on synonymous site evolution is maximal, and cooler, marginal regions of little selection. Small bacterial genomes (such as *Helicobacter pylori* or *Borrelia burgdorferi*) and large eukaryotic genomes (like those of *H. sapiens* or *M. musculus*) fall in these marginal regions, while the yeast genomes fall within the hot spot.

4.2 Genome size and genomic tRNA content as determinants of selection on codon usage

Perhaps the most important contribution of this work to the understanding of codon usage is the unexpected finding that genome size and tRNA gene redundancy define a genomic landscape where the action of natural selection on codon usage can be mapped to eukaryotic and prokaryotic organisms. Our findings suggest that an optimal combination of these factors exists, for which the action of translational selection is maximal. We can use the surface depicted in figure 4.7 to tentatively predict whether translational selection is operative in any given genome given that we know its size and the number of tRNA genes it contains. The challenge now is to identify the underlying causes that might explain this pattern.

The adaptation of codon usage to the genomic tRNA gene pool is a well known phenomenon in various organisms where translational selection is known to be present. In fact, some authors have discussed how the redundancy in the gene num-

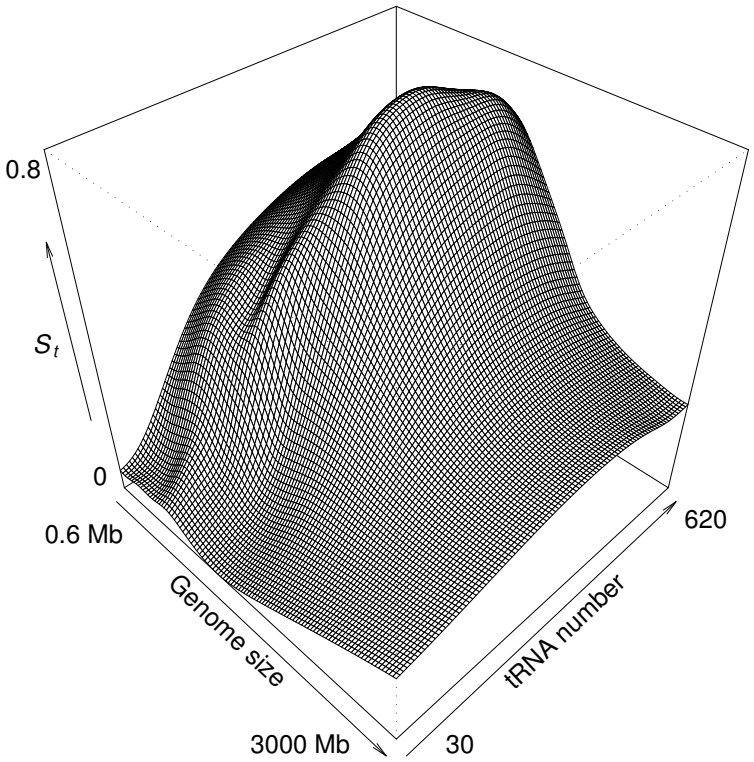

Figure 4.6: Perspective plot of the Gaussian process prediction on S_t values. The horizontal axes are in logarithmic scale. Although the smoothness of the surface is essentially derived from the data, it can be influenced by setting a prior on the scale parameter of the regression model. A prior that encouraged a smooth appearance of the regression surface was intentionally chosen. The MCMC GP model was implemented using the Flexible Bayesian Modelling Program Suite (R. Neil, http://www.cs.toronto.edu/~radford/fbm.software.html).

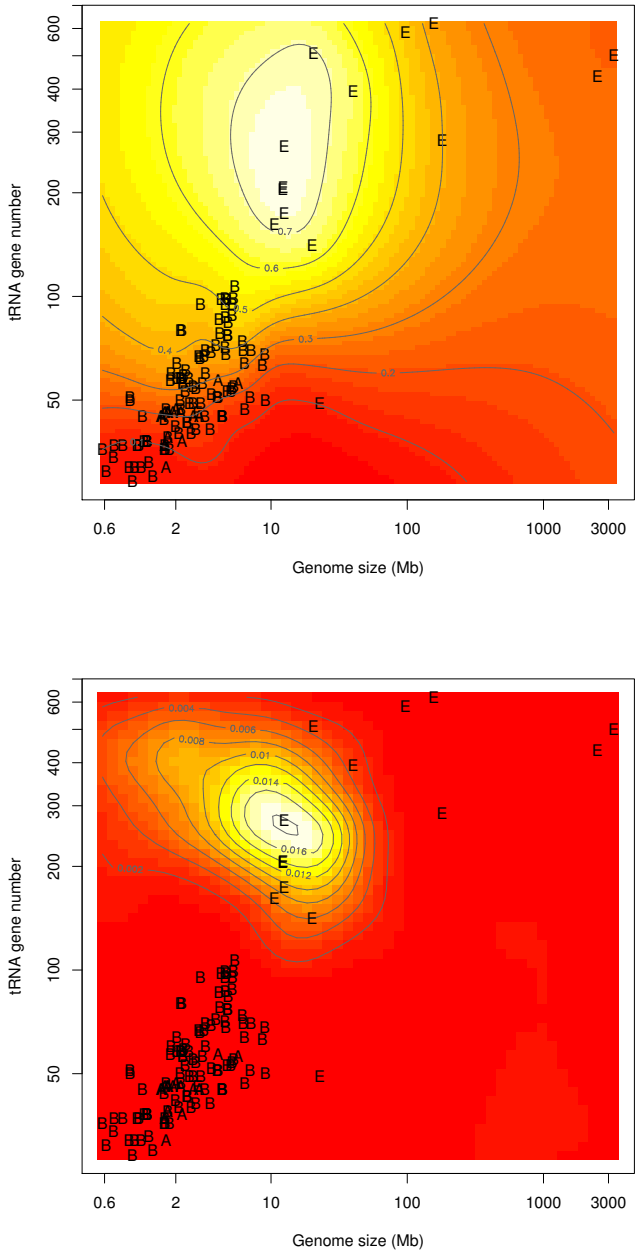

Figure 4.7: Thermal image representation of codon usage-tRNA coadaptation in eukaryotic and prokaryotic genomes.
Top panel: estimated S_t surface as a function of genome size and tRNA number. High, or hot regions are represented in white, while lower, or cooler zones are represented in red. Bottom panel: An advantage of the Bayesian approach is that parameters of interest can be given a probabilistic interpretation. For example, we were interested in locating the point of maximum translational selection in the regression surface. A density plot of the most likely location of this point can be easily obtained from the MCMC samples of surfaces by evaluating their maxima. The Gaussian surface was fitted using the Flexible Bayesian Modelling program suite (R. Neil, http://www.cs.toronto.edu/~radford/fbm.software.html). E: Eukaryotes, B: Bacteria and A: Archaea.

ber of certain tRNA isoacceptors matches the frequencies of the preferred set of codons in yeast and worm [33, 112]. However, what does not appear to have been discussed so far, is how the lack of duplicated tRNA genes might explain the absence of translationally selected codons in bacteria with small genomes. A vivid example of this is presented in the genome of *Helicobacter pylori*, where the absence of translationally selected codons is well documented [83]; *H. pylori* presents only 36 tRNA genes with only one tRNA species presenting two copies, it is this lack of tRNA gene duplication that determines the absence of translational selection in that organism. It can be argued that it is the need for translational optimisation and hence codon usage that shapes the tRNA pool of organisms [12]. However, we contend that selection favouring small genome size implies an overall reduction of the redundancy of the whole genome, that is, reduction of duplicated genes of any kind (tRNA, rRNA, protein genes, repetitive elements, etc.), and it is this kind of selective force, not codon usage itself, that shapes tRNA redundancy.

Since genome size and genomic tRNA number are highly correlated ($R = 0.84$), it seems logical to think that both factors coevolve in a concerted way. The evolution of genome size has largely been related to the evolution of repetitive DNA [114], so the mechanisms that explain the increase in copy number of selfish genes within genomes might be taken into account to partially explain the evolution of tRNA genes. In fact, the association between tRNAs and transposable elements is well documented in eukaryotic genomes [87, 140], and the evolution of tRNA genes themselves have been described as a repetitive process [66]. Note that we are not saying here that genome size determines tRNA gene number, but rather that there might be common mechanisms that explain the evolution of both. For example, imagine an horizontal transfer event in certain bacterial genome, the newly acquired chunk of DNA could contain several tRNA genes. This simple evolutionary event would both increase genome size *and* tRNA gene number. Of course, we could also have tandem replication of tRNA operons, and this would significantly increase ge-

nomic tRNA gene content, but not genome size. The forces that shape tRNA evolution need to be further investigated in order to understand codon usage evolution. It has recently been shown that genomic tRNA content correlates positively with bacteria growth rate [120]. What is puzzling though, is that genome size and growth rate seem to be uncorrelated in bacterial genomes. Furthermore, genome size can explain part of the variation in tRNA numbers after correcting for growth rate (data not shown). It seems that at least in bacterial genomes selection for growth rate has had an important impact on the evolution of tRNA redundancy.

The case for Eukaryotes seems to be more complicated. Figure 4.7 indicates that the larger eukaryotic genomes show low S_t values despite the fact that these genomes present the most redundant set of tRNA genes. So we cannot call upon translational selection to explain the variation observed in genomic tRNA sets in Eukaryotes. Lynch and Connery [92] have proposed that the evolution of genome complexity is linked to a reduction of effective population size throughout the living kingdoms. They argue that a reduction in effective population number is due to ecological constraints imposed on organisms with large body sizes. This, in turn, allows the maintenance of genomic features that would have otherwise been eliminated by purifying selection, such as: long introns, large multigene families, repetitive element dispersion, and expansion of genome size. If these ideas are proved correct, then they could explain nicely the features observed in the eukaryotic region of figure 4.7. We shall discuss tRNA evolution in more detail in the next chapter, and we will consider the ideas about genome complexity, and their implications on the evolution of codon usage in the final chapter of this dissertation.

4.3 Empirical correlation between S_t and the population parameter S

Sharp *et al*. [124] performed an extensive analysis of codon usage in 80 eubacterial genomes. These workers estimated the population parameter S according to the technique described in chapter 2, and compared their results to the S_t values obtained here for the 65 genomes common to both analyses. They found a significant correlation ($R = 0.46$) between both studies. It is important to note that in their work, Sharp *et al*., calculated S as an average over four optimal codons (those coding for Phe, Tyr, Ile and Asp) for a set of 40 proteins (mostly ribosomal proteins) that are expected to be expressed constitutively at high levels. On the other hand, S_t as defined in this work, takes into account all synonymous codons, and its value relates to the whole genome. For example, if a genome presents a very small set of highly expressed genes with substantial codon bias due to selection, but the rest of the genome's codon usage is driven mainly by mutation/drift, then S_t would be expected to be small, while S for the highly expressed gene set would be expected to be high. This represents a source of discrepancy between both codon measures (and this is well exemplified by the genome of *Bacillus subtitilis* where $S = 1.36$ and $S_t = -0.068$). Figure 4.8 shows a comparison between Sharp's S and a recalculated version of S_t that only considers genes related to translation and biogenesis (group J, clusters of orthologous groups [134]). This version of S_t indeed correlates highly with S ($R = 0.71$ after excluding *Nitrosomanas europaea* and *Xylella fastidiosa*, see figure 4.8 legend). It is very interesting that these rather different ways of analysing selection on codon usage share such striking similarities. Overall this shows that both analyses are highly complementary: while S simply measures the relative deviation in codon frequencies in two gene sets (deviations that could result, for example, from differences in mutation pattern in both sets as in the case of the two outlier genomes mentioned above), S_t confirms that the observed devi-

ations are due to adaptation to the genomic tRNA set, hence, reinforcing the idea that selection for translational optimisation has shaped the patterns of codon usage in highly expressed genes in the genome in question.

4.4 Selection on codon usage and growth rate in bacteria

It is a well acknowledged fact that fast growing microorganisms show a stronger selected codon usage bias [8]. However, because estimates of selection strength for a large sample of bacterial genomes were not previously available, it was not possible to study this relationship in detail. Fairly recently, Rocha [120] compiled a table of optimal duplication times for 101 bacterial genomes. This allows to test (perhaps for the first time) whether fast growing microorganisms are those which indeed show the strongest selected codon usage bias. Comparing the recalculated S_t values with Rocha's values shows that these two values correlate negatively (figure 4.9). Very similar results are obtained substituting Sharp's S values for S_t (data not shown).

4.5 Criticism and limitations of the model

The statistical model presented in this study failed to explain around 40% of the variation observed in S_t values in the genomes studied. The reasons for this are many, and include some of a technical, and others of a biological nature. The technical limitations are statistical in their character, and relate to the choice of regression model and the unusual structure of the data. tRNA gene numbers and genome size are highly correlated, so ample areas of genomic landscape do not present data points, therefore the predictions of the model in the upper left, and lower right corner of figure 4.7 are extrapolations that might not necessarily reflect the true be-

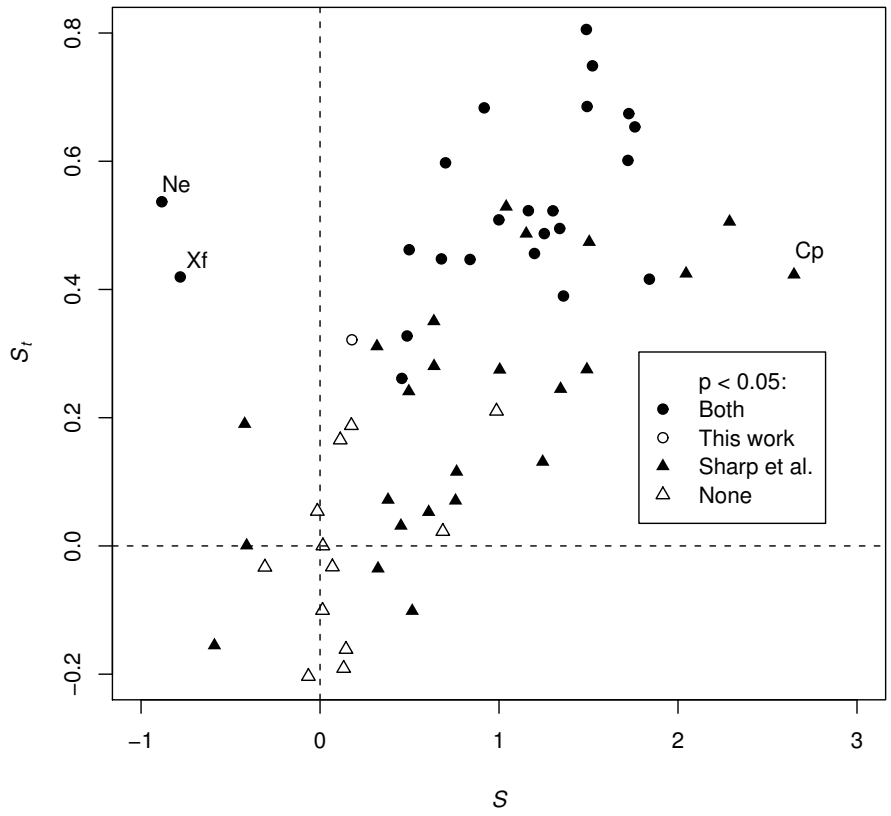

Figure 4.8: Comparison between estimates of S_t and S in prokaryotic organisms. S values from Sharp *et al*. [124]. S_t values where recalculated taking into account only genes belonging to group J according to the cluster of orthologous groups classification [134]. This group mainly contains ribosomal proteins, elongation factors, and other proteins related to biogenesis and translation. *Nitrosomanas europaea* (Ne) and *Xylella fastidiosa* (Xf) are the two most conspicuous outliers on the left hand side of the plot. These genomes present peculiar base compositions that bias their estimated S values. See [124] for a more detailed discussion. *Clostridium perfringens* (Cp) presents the highest S value but only a moderate S_t value. This is a very AT rich genome (14% GC3s), and the assumption that $\omega \approx 0$ (equation 3.11 on page 55) might not hold, giving way to an inaccurate estimate of S_t.

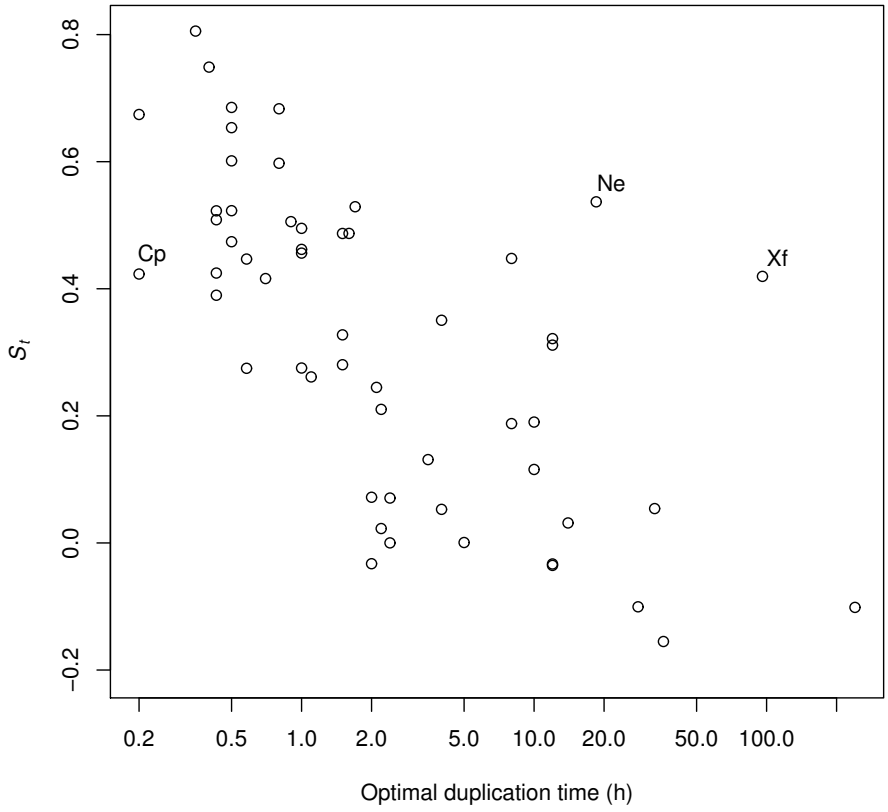

Figure 4.9: Selection on codon usage and growth rate in bacteria.
S_t values for orthologous group J, vs. optimal duplication time $(1/\mu)$ [120]. The correlation between the PIC for S_t and the optimal duplication time is $R = 0.67$. Cp: *C. perfringens*, Ne: *N. europeacea* and Xf: *X. fastidiosa*.

haviour of organisms in these regions. Another unusual characteristic of the data is the excessive oversampling of small genomes, which are easier and cheaper to sequence; this certainly affects the model in the sense that densely packed areas of data contribute more heavily to the shape of the regression surface than areas with more scattered data.

The choice of a particular regression analysis is also a problem. Classical parametric analyses like polynomial regressions tend to over-fit the data as the degree of the polynomial is increased [138]; and they also tend to produce surfaces that vary wildly in areas where the data is poorly represented. On the other hand, Gaussian processes are robust against highly correlated predictors, gaps in the data, or biased samples [118]. The final regression surface seen in figure 4.6 is an average of all the regression surfaces that explain the data equally well, so the wild fluctuations of classical regression surfaces over the areas where the data is poorly represented is averaged out. Furthermore, the inclusion of simulation and Bayesian analysis allows us to assess the validity of the regression surface, and obtain probability values for the features observed in it. If the model predicts a maximum in an area where the data is poorly represented, this maximum will have a very low probability because simulated surfaces in that area will tend to produce random maxima and minima; only regions that consistently produce the same feature are reliable. However, Gaussian processes also present limitations, this is a nonparametric model, so we lack a meaningful parametric equation to describe the data, and using it to predict accurate S values for new organisms with values outside the range of the current data is not necessarily appropriate.

The model also failed to account for some biological variables. For example secondary structure effects on codon preferences, or context dependent mutational biases were not explicitly taken into account (these are grouped together in the error term of equation 3.11). Another important variable that was considered in preliminary analysis was silent GC content, this variable can improve the predictive

power of the model to around 70% but its inclusion does not change the shape of the regression surface (data not shown). GC content and codon usage have been widely discussed in the literature, and it is thought that highly biased mutational patterns might restrict codon selection in certain organisms [91]. Some bacteria with extreme values of GC content, present low S_t values despite having intermediate genome sizes and tRNA numbers, examples of this are *Clostridium tetani* (GC = 0.29, $S_t = 0.029$), species of the genus *Streptomyces* (GC > 0.70 , $S_t < -0.101$) or *Borrelia burgdorferi* (GC = 0.29, $S_t = 0.115$) where asymmetrical replication is the major source of codon usage variation [99], and where the presence of translational selection has been debated [113]. This work also ignored the known fact that the proportion of tRNA isoacceptors can vary as a function of growth rate, tissue type, etc. and it has been suggested that genes accommodate their codon usage according to their particular tRNA environment [82]; in this work overall genomic tRNA content was used to calculate tAI and this might not always be appropriate. Also particular taxonomic groups or organisms with similar forms of life might show similar codon trends; for example, thermophilic bacteria have been shown to have the same codon preferences despite their large variations in overall GC content [93].

Perhaps the best way to exemplify the limitations of the model is by analysing the case of *Bacillus subtilis*. This organism presents an S_t value of -0.068, however, it has been reported that translational selection is operative in its genome [103]. Previous analysis on codon usage in this organism [81], showed that its genome can be divided into three gene classes according to their codon composition. Class II, which represents only 4.6% of open reading frames, comprises highly expressed genes that show biased codon patterns that can be explained through codon optimisation to match the tRNA pool of this organism [69]. Classes I (82.3%) and III (13.1%), which comprise the rest of the genome, show codon patterns determined by amelioration (nucleotide composition decay [106]) and mutation-random drift equilibrium; this contrasts with the genome of *Escherichia coli* K-12 where most

of the genome seems to be under translational selection [30]. It is evident that the selection effect on class II is hindered by classes I and III in our analysis when we consider the whole genome. Therefore, a small S_t value for a whole genome means that translational selection might be negligible at a genomic scale, but it can nonetheless have a strong effect on smaller scales, such as particular gene sets.

A final point is whether the trends observed for selected codon bias in Eukaryotes and Prokaryotes can be reconciled. Figures 4.3 and 4.4 show opposing relationships among S_t, tRNA numbers and genome size when each superkingdom is considered separately: while increasing tRNA numbers and increasing genome size correlate positively with selected codon bias in Prokaryotes, the opposite trend is actually observed in Eukaryotes. The maximum observed for S_t when all organisms are considered together (figure 4.5) is the transition point between the prokaryotic and eukaryotic superdomains. Is this a natural optimum for selected codon bias or an artifact of the regression model? After all there are no prokaryotic genomes beyond 10Mb and 150 tRNAs. I suspect the clue might lie with *Encephalitozoon cuniculi*. This Eukaryote has a tiny genome (2.9Mb), which is much smaller than the closest fungi genomes. Interestingly, *E. cuniculi* shows trends of codon bias that are very similar to those of bacteria with similar genome size and tRNA numbers. I suspect that as more genome sequences become available for intracellular parasitic Eukaryotes, a gradient in genome size and tRNA numbers between *E. cuniculi* and the yeast genomes will be observed, and the genome size gap between Eukaryotes and Prokaryotes will be filled. Whether this group of Eukaryotes would follow the same trends as some bacteria is an interesting question that remains open.

In the instances where selection has been associated with translational efficiency, both in Eukaryotes and Prokaryotes, the patterns of codon usage of highly expressed genes have been shown to mimic the structure of the genomic tRNA set [69, 68]. This suggests that the nature of the action of selection at the codon level in both taxa should be similar. Furthermore, the mathematical equations that describe the

population processes of codon usage in haploid and diploid organisms are essentially identical [13, 100]. Thus, it seems reasonable to treat Prokaryotes and Eukaryotes within the same framework when analysing codon usage. This does not mean that each group might not present its own idiosyncrasies relating to codon usage patterns. Prokaryotes and Eukaryotes differ widely in their lifestyles, and within these groups there is a large diversity of organisms adapted to substantially different environments and ecological niches. How a particular organism's lifestyle might determine its codon usage is a very interesting area of research that needs to be explored further. It is very likely that large parts of the variation observed in the regression model could be explained if variables related to lifestyle such as growth rate or optimal growth temperature were included. However one of the main ideas of this chapter is that there are indeed broad underlying patterns that are common to *both* eukaryotic and prokaryotic organisms. I believe that to understand the evolution of codon usage it is necessary to pay attention to the evolution of tRNA genes and the evolution of genome complexity. These issues will be explored in the following chapters.

Table 4.1: Codon usage tRNA coadaptation (S_t) in Prokaryotes and Eukaryotes.
Note: The asterisk indicates statistical significance ($p < 0.05$)

Kingdom	Organism	S_t	p		tRNA	Size (Mb)	Source
Archaea	*Aeropyrum pernix*	-0.172	0.865		46	1.7	NC_000854
	Archaeoglobus fulgidus	0.329	0.136		46	2.2	NC_000917
	Halobacterium sp.	-0.300	0.875		47	2.0	NC_002607
	Methanobacterium thermoautotrophicum	0.059	0.389		39	1.8	NC_000916
	Methanococcus jannaschii	0.133	0.341		37	1.7	NC_000909
	Methanopyrus kandleri	0.046	0.388		32	1.7	NC_003551
	Methanosarcina acetivorans	0.567	0.004	*	56	5.8	NC_003552
	Methanosarcina mazei	0.523	0.041	*	57	4.1	NC_003901
	Pyrobaculum aerophilum	0.024	0.447		38	2.2	NC_003364
	Pyrococcus abyssi	0.298	0.160		46	1.8	NC_000868
	Pyrococcus furiosus	0.447	0.032	*	46	1.9	NC_003413
	Pyrococcus horikoshii	0.380	0.097		46	1.7	NC_000961
	Sulfolobus solfataricus	0.504	0.008	*	45	3.0	NC_002754
	Sulfolobus tokodaii	0.502	0.002	*	45	2.7	NC_003106
	Thermoplasma acidophilum	0.299	0.138		45	1.6	NC_002578
	Thermoplasma volcanium	0.344	0.108		45	1.6	NC_002689
Bacteria	*Bacillus anthracis str. Ames*	0.443	0.018	*	95	5.2	NC_003997
	Bacillus cereus ATCC 14579	0.464	0.012	*	107	5.4	NC_004722
	Bacillus halodurans	0.090	0.321		78	4.2	NC_002570
	Bacillus subtilis	-0.068	0.575		86	4.2	NC_000964
	Bacteroides thetaiotaomicron VPI-5482	0.622	0.004	*	70	6.3	NC_004663
	Bifidobacterium longum	0.342	0.102		56	2.3	NC_004307
	Blochmannia floridanus	-0.052	0.570		37	0.7	NC_005061
	Bordetella bronchiseptica	-0.034	0.503		55	5.3	NC_002927

Kingdom	Organism	S_t	p		tRNA	Size (Mb)	Source
	Bordetella parapertussis	-0.023	0.545		53	4.8	NC_002928
	Bordetella pertussis	-0.188	0.796		51	4.1	NC_002929
	Borrelia burgdorferi	0.115	0.233		32	0.9	NC_001318
	Bradyrhizobium japonicum	-0.137	0.810		50	9.1	NC_004463
	Buchnera aphidicola	-0.016	0.530		31	0.6	NC_004545
	Campylobacter jejuni	0.045	0.408		44	1.6	NC_002163
	Caulobacter crescentus	0.034	0.415		51	4.0	NC_002696
	Chlamydia muridarum	-0.106	0.700		37	1.1	NC_002120
	Chlamydia trachomatis	-0.068	0.677		37	1.0	NC_000117
	Chlamydophila caviae	-0.057	0.616		38	1.2	NC_003361
	Chlamydophila pneumoniae AR39	-0.025	0.543		38	1.2	NC_002179
	Chlamydophila pneumoniae CWL029	-0.011	0.551		38	1.2	NC_000922
	Chlamydophila pneumoniae J138	-0.001	0.476		38	1.2	NC_002491
	Chlamydophila pneumoniae TW-183	-0.013	0.513		38	1.2	NC_005043
	Chlorobium tepidum	0.117	0.277		50	2.2	NC_002932
	Chromobacterium violaceum	0.235	0.181		98	4.8	NC_005085
	Clostridium acetobutylicum	0.192	0.105		72	3.9	NC_003030
	Clostridium perfringens	0.298	0.031	*	95	3.0	NC_003366
	Clostridium tetani	0.029	0.423		54	2.8	NC_004557
	Corynebacterium efficiens YS-314	0.473	0.041	*	56	3.1	NC_004369
	Corynebacterium glutamicum	0.701	0.007	*	60	3.3	NC_003450
	Coxiella burnetii	0.175	0.206		42	2.0	NC_002971
	Deinococcus radiodurans	0.100	0.297		49	3.1	NC_001263,NC_001264
	Enterococcus faecalis V583	0.337	0.100		67	3.2	NC_004668
	Escherichia coli CFT073	0.646	0.002	*	88	5.2	NC_004431
	Escherichia coli O157:H7 EDL933	0.654	0.005	*	99	5.3	NC_002655
	*Escherichia coli K12*l	0.747	0.002	*	87	4.6	U_00096

Kingdom	Organism	S_t	p		tRNA	Size (Mb)	Source
	Fusobacterium nucleatum	0.115	0.273		47	2.2	NC_003454
	Haemophilus ducreyi 35000HP	0.313	0.112		47	1.7	NC_002940
	Haemophilus influenzae	0.520	0.042	*	57	1.8	NC_000907
	Helicobacter hepaticus	0.021	0.462		36	1.8	NC_004917
	Helicobacter pylori 26695	-0.114	0.715		36	1.7	NC_000915
	Helicobacter pylori J99	-0.102	0.607		36	1.6	NC_000921
	Lactobacillus plantarum	0.297	0.114		70	3.3	NC_004567
	Lactococcus lactis	0.443	0.062		61	2.4	NC_002662
	Listeria innocua	0.298	0.166		66	3.0	NC_003212
	Listeria monocytogenes	0.282	0.101		67	2.9	NC_003210
	Mesorhizobium loti	0.024	0.496		51	7.0	NC_002678
	Mycobacterium bovis	-0.115	0.818		45	4.3	NC_002945
	Mycobacterium leprae	0.035	0.491		45	3.3	NC_002677
	Mycobacterium tuberculosis CDC1551	-0.092	0.752		45	4.4	NC_002755
	Mycobacterium tuberculosis H37RV	-0.099	0.711		45	4.4	NC_000962
	Mycoplasma gallisepticum R	0.210	0.165		32	1.0	NC_004829
	Mycoplasma genitalium	0.238	0.178		36	0.6	NC_000908
	Mycoplasma penetrans	0.291	0.085		30	1.4	NC_004432
	Mycoplasma pneumoniae	0.046	0.449		37	0.8	NC_000912
	Mycoplasma pulmonis	0.176	0.179		29	1.0	NC_002771
	Neisseria meningitidis MC58	0.727	0.005	*	59	2.3	NC_003112
	Neisseria meningitidis Z2491	0.721	0.000	*	58	2.2	NC_003116
	Nitrosomonas europaea	0.420	0.078		41	2.8	NC_004757
	Nostoc sp	0.330	0.054		47	6.4	NC_003272
	Oceanobacillus iheyensis	0.242	0.116		69	3.6	NC_004193
	Pasteurella multocida	0.485	0.011	*	56	2.3	NC_002663
	Pirellula sp	0.328	0.154		70	7.1	NC_005027

Kingdom	Organism	S_t	p		tRNA	Size (Mb)	Source
	Porphyromonas gingivalis W83	0.408	0.031	*	53	2.3	NC_002950
	Prochlorococcus marinus str. MIT 9313	0.079	0.374		39	1.8	NC_005071
	Prochlorococcus marinus subsp. CCMP1375	0.202	0.143		37	1.7	NC_005042
	Prochlorococcus marinus subsp. CCMP1986	0.226	0.141		43	2.4	NC_005072
	Pseudomonas putida KT2440	0.602	0.023	*	74	6.2	NC_002947
	Pseudomonas syringae	0.519	0.010	*	64	6.4	NC_004578
	Rickettsia conorii	0.093	0.269		33	1.3	NC_003103
	Rickettsia prowazekii	0.178	0.179		32	1.1	NC_000963
	Salmonella enterica Typhi Ty2	0.692	0.005	*	77	4.8	NC_003198
	Salmonella enterica Typhi	0.691	0.000	*	77	4.8	NC_004631
	Salmonella typhimurium LT2	0.672	0.002	*	84	4.9	NC_003197
	Shigella flexneri 2a-2457T	0.705	0.003	*	99	4.6	NC_004741
	Shigella flexneri 2a-301	0.638	0.000	*	95	4.6	NC_004337
	Sinorhizobium melitoti	0.209	0.163		52	3.7	NC_003047
	Staphylococcus epidermidis	0.347	0.010	*	58	2.5	NC_004461
	Streptococcus agalactiae 2603VR	0.370	0.058		80	2.2	NC_004116
	Streptococcus agalactiae NEM316	0.392	0.000	*	80	2.2	NC_004368
	Streptococcus mutans	0.242	0.159		64	2.0	NC_004350
	Streptococcus pneumoniae-R6	0.466	0.059		58	2.0	NC_003098
	Streptococcus pneumoniae-TIGR4	0.472	0.041	*	58	2.2	NC_003028
	Streptococcus pyogenes	0.459	0.038	*	60	1.9	NC_002737
	Streptomyces avermitis	-0.140	0.723		68	9.0	NC_003155
	Streptomyces coelicolor	-0.101	0.684		63	8.7	NC_003888
	Synechococcus sp. WH8102	0.173	0.302		43	2.4	NC_005070
	Synechocystis sp. PCC6803	0.427	0.075		41	3.6	NC_000911
	Thermoanaerobacter tengcongensis	0.302	0.102		55	2.7	NC_003869
	Thermosynechococcus elongatus	0.339	0.071		40	2.6	NC_004113

Kingdom	Organism	S_t	p		tRNA	Size (Mb)	Source
	Treponema pallidum	0.205	0.153		45	1.1	NC_000919
	Tropheryma whipplei Twist	0.252	0.149		50	0.9	NC_004572
	Tropheryma whipplei TW0827	0.213	0.282		51	0.9	NC_004551
	Vibrio cholerae	0.074	0.372		98	4.3	NC_002505,NC_002506
	Wigglesworthia brevipalpis	0.037	0.403		34	0.7	NC_004344
	Wolinella succinogenes	0.208	0.169		40	2.1	NC_005090
	Xanthomonas campestris pv. campestris str. ATCC 33913	0.269	0.133		53	5.1	NC_003902
	Xanthomonas axonopodis pv. citri str. 306	0.357	0.126		54	5.2	NC_003919
	Xylella fastidiosa 9a5c	0.315	0.043	*	49	2.7	NC_002488
	Xylella fastidiosa Temecula1	0.396	0.044	*	49	2.5	NC_004556
	Yersinia pestis CO92	0.486	0.013	*	68	4.7	NC_003143
	Yersinia pestis KIM	0.537	0.003	*	71	4.6	NC_004088
Eukarya	*Arabidopsis thaliana*	0.266	0.168		620	157.0	
	Caenorhabditis elegans	0.460	0.017	*	585	97.0	
	Candida glabrata	0.856	0.000	*	207	12.3	
	Cryptococcus neoformans	0.864	0.000	*	141	20.0	
	Debaryomyces hansenii	0.747	0.001	*	205	12.2	
	Drosophila melanogaster	0.331	0.068		285	180.0	
	Encephalitozoon cuniculi	0.180	0.160		46	2.9	
	Homo sapiens	0.211	0.114		501	3289.0	
	Kluyveromyces lactis	0.808	0.000	*	162	10.6	
	Mus musculus	0.297	0.086		435	2493.0	
	Neurospora crassa	0.674	0.000	*	394	40.0	
	Plasmodium falciparum	0.129	0.233		49	22.9	
	Saccaromyces cerevisiae	0.777	0.001	*	273	12.5	
	Schizosaccaromyces pombe	0.848	0.000	*	174	12.5	
	Yarrowia lipolytica	0.814	0.000	*	510	20.5	

5 Transfer RNA evolution and codon usage

As discussed previously, selection on codon usage and the structure of genomic tRNA sets seem to be closely related. Every organism presets a particular configuration of tRNA genes in its genome. Some tRNA gene species might be found in multiple copies while others are present in a single one [96, 48]. The particular configuration of tRNAs and their corresponding anticodons can be defined as the tRNA anticodon system of a particular genome. Thus, when the tRNA anticodon composition matches the codon usage of protein coding genes, we speak of codon usage optimisation due to natural selection. Understanding why and when both preference systems match in a given genome is important for understanding codon evolution.

Multivariate analysis shows that eukaryotic and prokaryotic genomes can be easily distinguished according to their tRNA anticodon preferences (figure 5.1). Yeast has a much more similar tRNA structure to human than to *E. coli*, despite the fact that that both yeast and *E. coli* present highly optimised codon usage. This is not surprising since we can assume that the human and yeast tRNA pool represent an expansion of some ancestral eukaryotic pool, and all modern eukaryotic tRNA pools thus resemble this ancestral configuration. Eukaryotic genomes have high numbers of tRNA genes with ANN anticodons (INN after post-transcriptional modification) that recognise codons ending in T or C (and to a smaller degree A) [96, 48]. On the other hand, bacterial genomes tend to use tRNAs with GNN anticodons that recog-

nise the corresponding codons ending in T or C [96]. The use of ANN-tRNAs is the chief cause of the marked separation between Eukaryotes and Prokaryotes along the first axis in figure 5.1. The second axis is explained by variation in G+C content in the genomes studied (not shown). Thus, the question is to understand how the modern configuration of tRNA sets evolved and whether this evolutionary process affected codon usage. What evolutionary constraints have dictated the tRNA anticodon preferences of Eukaryotes and Prokaryotes? What are the mechanisms that dictate the expansion/contraction of tRNA sets? How have tRNA anticodon preferences evolved through particular phylogenies? Have idiosyncratic mutational patterns within genomes affected the structure of tRNA sets? Unfortunately, very little work has been carried out in this area. In order to fill this gap, this chapter presents a brief exploration of tRNA evolution and its relationship to codon usage, paying particular detail to the bacterium *Escherichia coli*.

5.1 Evolution of transfer RNA genes in *Escherichia coli*

The bacterium *Escherichia coli* has been one of the most widely studied organisms in molecular biology. At the time of writing (March 2007), the genome of at least eight strains and sub strains of this bacterium have been fully sequenced, and at least fourteen more are in progress or in the assembly phase[1]. This does not consider the genomes of the various strains of *Shigella flexneri* that have been sequenced, an organism that is now known to be a form of *E. coli* [117]. This makes *E. coli* well suited as a model organism to study evolution at the genomic level. This section explores the genomic structure and evolution of the tRNA clusters present in the genomes of *E. coli* strains K12 (*Ec*K12), O157:H7 EDL933 (*Ec*O157), CFT073 (*Ec*CFT073); *S. flexneri* 2a 301 (*Sf*2a301) and *Salmonella ty-*

[1] http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi

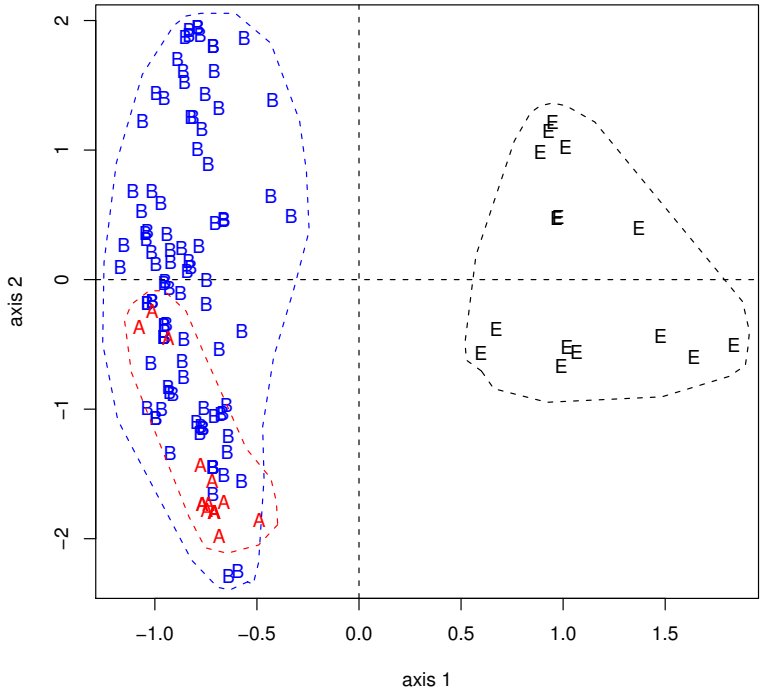

Figure 5.1: Correspondence analysis on tRNA anticodon preferences.
A matrix was constructed where each row represents a particular genome i and each column represents a particular anticodon j. Each cell contains the number of tRNA genes with anticodon j present in genome i. Correspondence analysis (figure 1.1, chapter 1) was applied to this matrix, and the two axis that explain the largest variation were obtained. Organisms are the same as in table 4.1 on page 81. E: Eukaryotes, B: Eubacteria, A: Archaea.

phimurium LT2 (*St*LT2). The availability of the genome sequence of these closely related organisms allows the easy identification of sets of orthologous tRNA genes, their rate of divergence, and the identification of duplication, insertion, deletion and inversion events during the evolutionary life of these genes. The analysis of the evolutionary dynamics of the genomic tRNA pool in these organisms might shed light onto the more general problem of tRNA evolution in other organisms and its relationship to patterns of codon usage bias.

5.1.1 Phylogenetic analysis

The five *E. coli* and *S. typhimurium* strains were aligned using MultiPipMaker [123] with *Ec*K12 as the reference organism. Genetic distances (estimated number of nucleotide substitutions per site) were initially computed from the alignment using Kimura's two parameter model (K2P, [74]). Assuming a constant molecular clock, and utilising the estimated K2P genetic distances a UPGMA tree was initially built [143] that shows the phylogenetic relationships for the genomes analysed (figure 5.2). The root of the tree was confirmed using *St*LT2 as an outgroup. The K2P distances are grossly underestimated since the naive K2P method fails to account for rate variation among sites [148]. To overcome this, and to obtain better estimates of divergence times, the branch lengths were optimised by maximum likelihood (ML) with PAML [149], utilising a general time reversible model of nucleotide substitution [146] and assuming rate variation among sites under a molecular clock (discrete gamma [148]). The UPGMA topology was later confirmed by an extensive ML search of the tree space. The divergence time between the *Escherichia* and *Salmonella* lineages has been estimated to be around 100 million years (Myr) [107, 27]. Taking into account this estimate, divergence times for each of the branches of the ML optimised tree were computed. Strains *Ec*K12, *Ec*O157 and *Sf*2a301 diverged roughly 6-7 Myr ago in a quasi star phylogeny, and this group diverged from *Ec*CFT073 around 11 Myr ago. The close relationship

of *Sf*2a301 to the *Escherichia* strains agrees with previous studies [117, 119] that identify *Shigella* spp. as a form of *E. coli*. However, the results presented here strongly disagree with previously estimated divergence times. The reasons for this discrepancy are discussed elsewhere [143].

It is important to note that bacterial genomes suffer extensive recombination and horizontal gene transfer among closely related strains (see [80] for a good review on this topic). This genetic flow implies different, conflicting gene genealogies along the bacterial chromosome and among strains. In this sense, the tree shown in figure 5.2 is an average over several distinct topologies. This tree is used as a framework onto which to study the evolution of tRNA genes in *E. coli*. This is simply a first approximation to this complicated problem. The most variable regions in the genomes studied were not used when building the tree, so it is hoped that the tree broadly reflects the more conserved, stable chromosomic backbone, rather the more mobile, less conserved elements [80]. However, further work is required to assess the validity of this assumption.

5.1.2 Genomic structure and evolution of tRNA genes

In order to identify the tRNA genes present in the *Escherichia-Shigella* clade, the five genomes were scanned with the computer package tRNAscan-SE [90]. Two low sensitivity programs analyse genomic sequence in order to identify candidate tRNA genes, these filtered sequences are then analysed by a highly selective tRNA covariance model [35]. The tRNAscan-SE predictions were compared for agreement with the annotated genomes. The genomes analysed present between 85 and 99 tRNA genes. These genes comprise between 45 and 52 different tRNA species representing between 40 and 41 anticodons (table 5.1). Here, two tRNA genes are considered to be of the same species if they present the same anticodon and if they present no more than 0.10 substituted sites (K2P) in pairwise comparisons. Most tRNA gene species are presented in one copy within the chromosome, with fewer

presented in two or more copies, with up to seven copies for the most abundant tRNA species in *Ec*O157. Accordingly, the distribution of tRNA species vs gene copy number is long-tailed (figure 5.3).

To identify the orthologous sets of tRNA genes in the *Escherichia-Salmonella* clade, distance matrices (K2P) were constructed to compare all tRNA genes between any two genomes. The resulting matrices[2] (kindly provided by M. Withers), gave a first approximation of the orthologous sets and their evolution. Discrepancies in determining detailed orthologous relationships were solved through genomic alignments of the problematic regions (pairwise Blast) and analysing marker genes and sequences present upstream and downstream of the tRNAs. Seventy-eight tRNA genes were identified in *Ec*K12 that have unambiguous orthologs in *St*LT2. Interestingly, these genes are highly conserved despite the fact that both organisms diverged about 100 Myr ago. Very few nucleotide substitutions were observed when the 78 genes in *Ec*K12 were aligned to their corresponding *St*LT2 orthologs, implying a low substitution rate (table 5.2). This contrasts sharply with the overall genomic substitution rate (table 5.2). Similar results are obtained when a full comparison among all the strains is carried out. Figure 5.4 depicts a detailed syntenic map showing the identified set of orthologous tRNA genes across all the strains analysed. In general, a well conserved tRNA backbone, that reflects the tRNA configuration of the last common ancestor of the *Escherichia-Shigella* clade, is present in all strains. The ancestral set of tRNA genes for the *Escherichia-Shigella* clade was determined from the syntenic map and the phylogenetic tree obtained above. This set was reconstructed as to minimise the number of evolutionary events along the tree branches. This ancestral set represents a tRNA core that is incredibly well conserved among the genomes analysed despite 100 Myr of evolution. With the exception of tRNA-rRNA operons, polycistronic tRNAs (figure 5.4) form part of this core and are well conserved across all strains including *St*LT2,

[2]The rather large matrices are available online at http://people.cryst.bbk.ac.uk/~fdosr01/trnas/

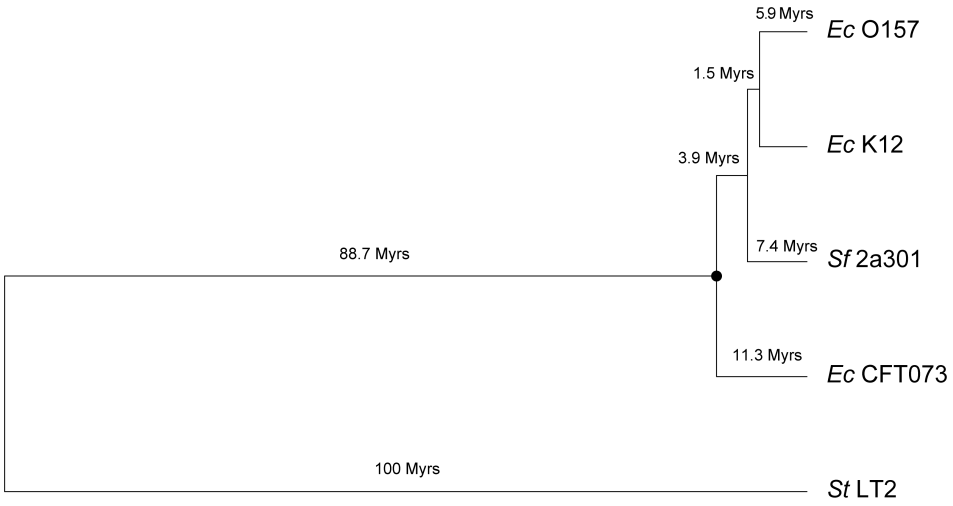

Figure 5.2: UPGMA tree with ML optimised branches for the *Escherichia-Shigella* clade.
Divergence times are given in million years (Myr). The black circle indicates the point of the ancestral tRNA reconstruction. Branch lengths were estimated taking into account rate variation among sites [148].

Table 5.1: Number of tRNA genes, tRNA species and number of anticodons present in the *Escherichia-Salmonella* clade.

	tRNA genes	tRNA species	Anticodons
*Ec*K12	86	45	40
*Ec*O157	99	52	41
*Sf*2a301	97	52	40
*Ec*CFT073	88	48	41
*St*LT2	85	47	40

Table 5.2: Substitution rates between *Escherichia coli* and *Salmonella typhimurium.*

Type	Rate (per site per Myr)
Overall genome	1.2×10^{-9}
tRNA genes	
Functional tRNA genes	1.4×10^{-11}
Protein coding genes	
Synonymous[1]	1.57×10^{-8}
Non synonymous[1]	3.7×10^{-10}

[1]Maximum likelihood estimate of codon substitution rates [150] for 2631 orthologous genes in the *Escherichia-Shigella* clade.

suggesting they were already present in their modern configuration more than 100 Myr ago in the *Escherichia-Salmonella* ancestor.

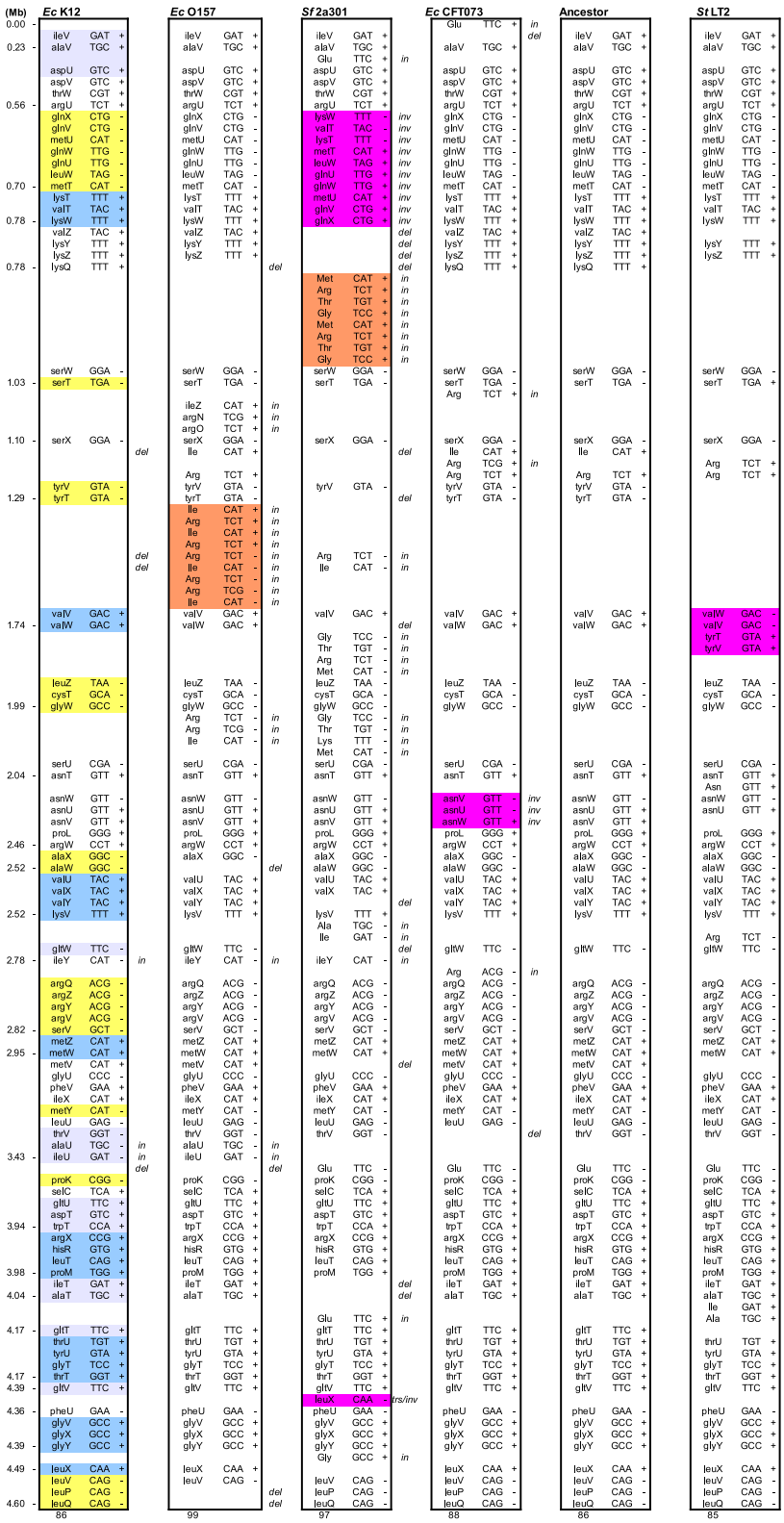

Figure 5.4: Orthologous tRNA gene sets for the *Escherichia-Salmonella* clade. (Full legend on next page).

(Figure 5.4 from previous page) Blue and yellow boxes show polycistronic tRNA clusters in *Ec*K12, grey boxes show tRNA-rRNA operons [63]; purple boxes show inverted regions when compared to the ancestral reconstruction; orange boxes show regions with intense tRNA insertions. The numbers on the left hand side indicate the genomic coordinates of tRNA genes in *Ec*K12. The numbers at the bottom indicate the total number of functional tRNA genes in each genome. The plus (+) and minus (-) signs indicate weather the gene is present in the leading or lagging strand of the chromosome. in: inserted, del: deleted, inv: inverted and trs: translocated.

Although tRNA genes are well conserved, their position and number of copies within the chromosome is not. Seventy-six evolutionary events can be identified along the tree branches of the *Escherichia-Shigella* clade, that include 49 insertions, 23 deletions, three inversions and one translocation (table 5.3). The most active genome seems to be *Sf*2a301 where the highest number of evolutionary events was detected, followed by that of *Ec*O157. These two genomes contain large numbers of insertions that are probably related to horizontal transfer events. *Ec*K12 and *Ec*CFT073 presented the best conserved tRNA sets.

The approach applied in this work does not allow the identification of tRNA evolutionary events in *St*LT2, since this organism was used as an outgroup, and ancestral character reconstruction at the most ancient node in the tree is uncertain [26]. Transfer RNA gene *gltT* can illustrate this point. This gene is present in *Ec*K12, *Ec*O157, *Ec*CFT073, and *Sf*2a301, but it is absent in *St*LT2 (figure 5.4). Two scenarios are possible: (1) *gltT* was not present in the *Escherichia-Salmonella* ancestor, and it was inserted in the genome of the *Escherichia* ancestor after divergence from *St*LT2; (2) *gltT* was indeed present in the last *Escherichia-Salmonella* ancestor, but it was deleted somewhere in the *Salmonella* lineage. Both explanations seem equally likely, because both involve only one evolutionary event. With current data it is impossible to distinguish the most likely explanation. However, the resemblance of *St*LT2 to the ancestral reconstruction suggests that its tRNA set has remained reasonably conserved.

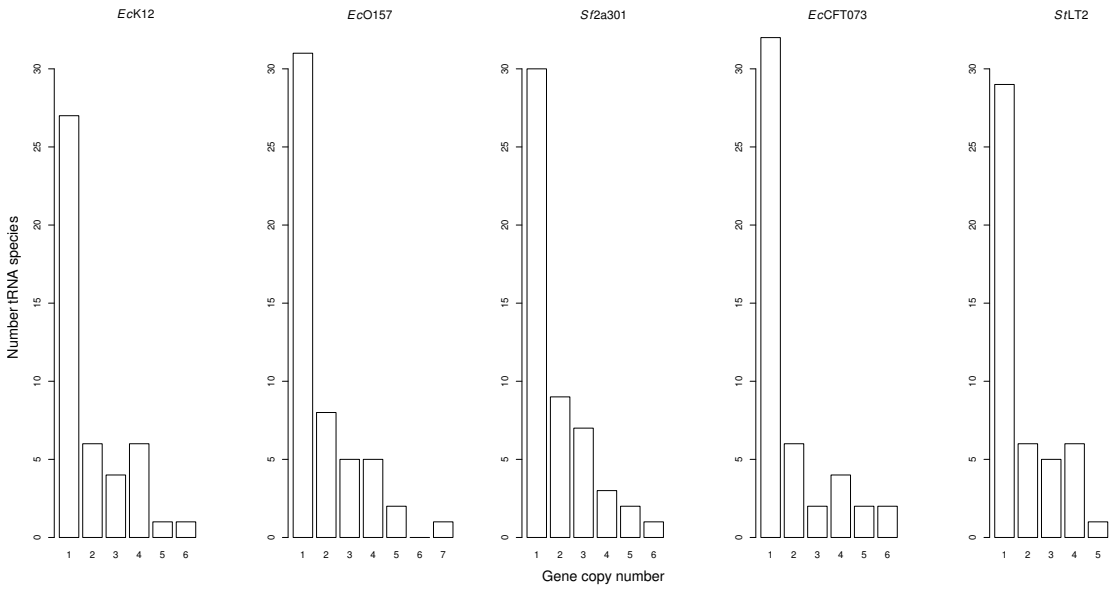

Figure 5.3: Distribution of number of tRNA species vs gene copy number for the *Escherichia-Shigella* clade.

Table 5.3: Estimated number of tRNA gene evolutionary events along the *E. coli* clade.

	*Ec*K12	*Ec*O157	*Sf*2a301	*Ec*CFT073	Per lineage	Per ling-Myr
Insertions[1]	3	18	24	4	12.25	1.08
Deletions	4	5	12	2	5.75	0.51
Inversions	0	0	2	1	0.75	0.07
Translocations	0	0	1	0	0.25	0.02

[1] and duplications, since both are indistinguishable under the current context.

5.1.3 Pseudo tRNAs

Nine putative pseudo tRNAs are apparent in the *Escherichia-Shigella* clade (table 5.4). These genes can be distinguished from normal tRNA genes by their unusually low covariance model score [35, 90] (figure 5.5). Six of these sequences could be confirmed as real pseudogenes since they have clear paralogs across the respective genomes and present high nucleotide substitution rates (table 5.4). For example, *thrW* probably underwent a duplication event 13-19 Mys ago, in the lineage leading to the *Ec*K12-*Ec*O157-*Sf*2a301 clade. One of the duplicates became nonfunctional giving rise to the pseudogene. A further duplication event occurred later on in the branch leading to *Sf*2a301 with similar results. These four pseudogenes are not full length copies, so they might have been used as points of recombination/horizontal transfer events in the genomes in question. The remaining three sequences identified by tRNAscan-SE as pseudogenes are highly conserved with hardly any substitutions observed for at least the last 19 Myr. These three orthologs do not show significant similarity to other tRNAs or to other sequences present in GenBank. These are perhaps truly functional RNAs awaiting discovery.

tRNAscan-SE identified only one putative pseudo tRNA in the *St*LT2 genome (table 5.4). This gene seems to be a paralog of *argU*. It is not clear whether *argU* underwent a duplication more than 100 Myr before the Escherichia lineage diverged from Salmonella, with the subsequent loss of the pseudogene in *Ec*O157, *Sf*2a301 and *Ec*CFT073, or whether two duplication events occurred independently in the *Ec*K12 and the *St*LT2 lineage. The relatively low divergence between the *St*LT2 pseudogene and its functional *Ec*K12 *argU* paralog (table 5.4) strongly suggests this pseudogene originated in a more or less recent and independent duplication event in the *St*LT2 lineage. The inclusion of other Salmonella strains into the phylogenetic tree might be needed in order to solve the issue.

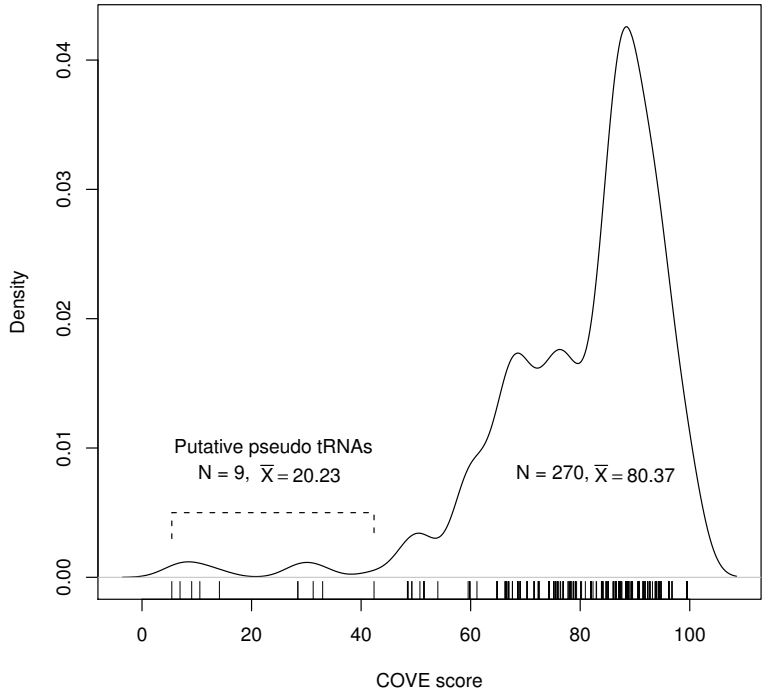

Figure 5.5: Kernel density estimate of COVE scores for all tRNA genes identified by tRNAscan-SE in the *Escherichia-Shigella* clade.

Transfer RNA and tRNA-like genes were identified by running tRNAscan-SE with sensitive settings (COVE cutoff score 1). A low COVE score simply means that the sequence in question shows distant similarity to the covariance tRNA model used. This could mean that the tRNA is nonfunctional (pseudo-tRNA), that it is a special type of tRNA (viral, organellal, etc.), or that it is another sort of regulatory RNA or non expressed, RNA-like sequence present in the genome [35].

5.1.4 The *Escherichia-Shigella* genomes present a well conserved chromosomic tRNA backbone

Considering the results presented in this chapter, tRNA evolution can be divided in three categories: (i) Evolution of tRNA sequences themselves; (ii) evolution of polycistronic tRNAs and the tRNA backbone; and (iii) evolution of interspersed tRNA genes. The first category relates to the most fundamental mode of evolution, with the slow accumulation of nucleotide substitutions, slow divergence, and reflecting the action of purifying selection on tRNA structure. The other two categories are related to the genomic organisation of tRNA genes, and probably reflect broader patterns of chromosome evolution in bacteria, like recombination and horizontal gene transfer.

Transfer RNA sequences evolve slowly, with tRNA orthologs in *Escherichia* spp. and *Shigella* spp. being nearly identical despite the fact that both groups diverged more than 100 Myr ago [9]. The estimated substitution rate in functional tRNA genes is several times smaller than the average genomic rate, the rate for synonymous sites and the average for non-synonymous sites in protein coding genes (table 5.2). The genomic rate is an average that includes substitutions in coding and non-coding regions, and it reflects a balance between regions under strong selection and regions that accumulate substitutions freely. The genomic rate is hence a good baseline to which comparisons can be made. Synonymous substitutions are those that happen inside protein coding genes but that do not change the encoded amino acid, these substitutions are relatively free from selective pressures. On the other hand, non synonymous mutations do change the encoded amino acid and hence are usually subjected to purifying selection [126]. The fact that tRNA genes show such low substitution rates compared to non synonymous substitutions, indicates that strong purifying selection acts on them in order to conserve their structure and identity [98].

The ancestral tRNA set contained 86 tRNA genes. Seventy eight genes from

this set are still present, and have remained nearly intact, in the lineages leading to *Ec*K12 and *St*LT2. These 78 genes reflect a conserved tRNA backbone that has remained well conserved for more than 100 Myr, and that can be easily distinguished in all the genomes analysed. It is clear that strong selective pressure must be operating to maintain such a high degree of conservation after 100 Myr of divergence. It should be noticed that the conserved backbone is made up mostly of polycistronic tRNAs, and mixed (tRNA and protein) operons [63]. How this set originated is an interesting question that needs to be addressed. A cluster dispersion model of tRNA evolution has been proposed to account for the generation of new tRNA species by duplication and coadaptation to the genetic code in the ancient lineages of life [145]. An extension of this model could perhaps be used to understand the modern tRNA structure of Eubacteria; however, a more extensive analysis, including more distantly related bacteria, is needed in order to work out ancient evolutionary events that led to the formation of the modern tRNA clusters seen in *Escherichia* spp. and *Salmonella* spp.

Most evolutionary events observed in the genomes analysed have involved interspersed tRNA genes. The genomes with the largest amounts of horizontally transferred DNA (*Ec*O157 and *Sf*2a301 [117, 56]) showed the largest number of tRNA insertions. It has been suggested that tRNAs that code for codons that are rare in the bacterium chromosome may be needed to express some of the foreign protein genes [56]. By comparison, *Ec*K12, which shows the smallest genome and the best conserved tRNA set, also presented the smallest number of tRNA related events.

5.1.5 Coevolution of codon usage and transfer RNAs

Since tRNA expression levels and tRNA gene copy number are correlated [61, 10, 69], it would be interesting to investigate whether the variation in gene copy number observed for some tRNA genes along the lineages analysed has had an effect on the codon usage of the genomes involved. In order to achieve this goal the S_t

values (chapter 3) for the coadaptation between genomic tRNAs and codon usage were computed for each genome, and the process was carried out using modern and ancestral tRNA sets. Table 5.5 shows the estimated S_t values for the members of the *Escherichia-Shigella* clade. It can be seen that, with the exception of *Ec*K12, the S_t values estimated from the ancestral reconstruction are higher than those obtained from modern tRNAs. This indicates that the codon usage of the modern genomes is better adapted to the reduced conserved core of tRNAs than to the complete modern set.

Bulmer [12] developed an elegant mathematical model relating the evolution of codon usage with tRNA abundance within the cell. In his model codon usage and tRNA genes coevolve in a feedback manner: the most abundant tRNAs drive up the frequencies of their cognate codons, and certain codons already present in high frequencies can in turn drive up the intracellular levels of their respective cognate tRNAs. A critical point of this model is that mutations that produce small changes in the expression level of some tRNA genes might be positively selected. In this case, the mutant tRNA presents a modified relative expression level that matches more closely the frequency of the respective codon. Although developed more than 19 years ago, this model does not seem to have empirical confirmation yet.

Transfer RNA expression levels within the cell are regulated by promoter sequences and tRNA gene copy number [61, 82, 63, 69]. Transfer RNA sequences presented in multiple copies within the bacterial chromosome tend to be expressed at higher levels than sequences with single copies. The number of copies of each tRNA species has been shown to be linearly and positively correlated to their respective expression levels. Bulmer's model does not take into account the dramatic changes in expression levels that might be brought about with the insertion or deletion of tRNA sequences. It is also unclear how mutations in the promoter regions of polycistronic tRNAs might affect the model, since some of these clusters contain different tRNA species, and a mutation in the promoter would affect the expression

Table 5.4: Putative pseudo tRNA genes identified by tRNAscan-SE in the *Escherichia-Shigella* clade.

Genome	Type	Start	End	COVE	Paralog
*Ec*K12	Thr	296402	296478	42.36	*thrW*
	Undet	344558	344638	28.45	- *
	Arg	585242	585324	9.07	*argU*
*Ec*O157	Thr	310574	310644	6.94	*thrW*
	Undet	402035	402115	31.25	- *
*Sf*2a301	Thr	328165	328094	10.56	*thrW*
	Thr	317706	317620	5.44	*thrW*
*Ec*CFT073	Undet	434982	435062	28.45	- *
	Thr	1211231	1211306	14.13	*thrU*
*St*LT2	Arg	617984	618056	20.46	*argU*

The start and end columns are the genomic coordinates for each pseudo gene. (*) These three RNAs are considered orthologous.

Table 5.5: S_t values for modern and ancestral tRNA sets.

	Modern	Ancestral	Difference	95%
*Ec*K12	0.747	0.723	0.0235	(-0.00577, 0.00526)
*Ec*O157	0.654	0.726	-0.0722	(-0.0376, 0.0370)
*Sf*2a301	0.638	0.684	-0.0457	(-0.00695, 0.00759)
*Ec*CFT073	0.646	0.696	-0.0500	(-0.0102, 0.0105)

Modern and ancestral w-values (equation 3.8) needed to obtain S_t were computed using the modern tRNA sets and the ancestral tRNA reconstruction respectively. S_t values were calculated only for genes with more than 100 codons. A randomisation test was performed to assess the statistical significance of the difference in modern and ancestral S_t values at 5% confidence. Two sets of tAI values were calculated for each gene using modern and ancestral w-values. Both sets were mixed and two new sets of tAI values were extracted randomly, new pseudo S_t values were obtained for both sets and the difference recorded. The procedure was repeated 1,000 times to obtain the 95% acceptance intervals shown.

levels of all these tRNA species simultaneously.

For the majority of genomes analysed, present codon usage frequencies match more closely the conserved core of tRNA genes than the full modern set. This might have important implications for the understanding of codon usage and tRNA coevolution. The large number of tRNA insertions and deletions observed during the evolutionary history of *Escherichia* spp. might have had profound effects on the relative expression levels of the different tRNA species. It is interesting then, that modern codon usage is finely tuned to the ancestral, well conserved, tRNA backbone. Our data suggests that this conserved backbone arose more than 100 Myr ago, possibly the product of a series of tRNA duplications, deletions and rearrangements. Whether this backbone represents a frozen accident during the evolutionary history of the *Escherichia* lineage that shaped its codon usage is an important question that needs to be addressed. It is possible that minor tuning between promoter regulated tRNA expression and codon usage happened later on in the evolutionary history of these organisms [63], but with tRNA gene copy number being responsible for the broader patter of codon usage seen in *Escherichia coli*. It is also unclear why the additional tRNA genes that have been inserted into the genomes of *Ec*O157 and *Sf*2a301 have not had an effect on codon usage trends in these organisms. Why and when the tRNA genomic system is decoupled from the codon usage is an interesting question that needs to be addressed. A detailed analysis of codon substitution patterns in protein coding genes for the whole *E.coli* genome and their relationship to tRNA evolution is needed in order to clarify this issue. This is the topic of the next chapter.

5.2 Evolution of tRNA genes as a repetitive process

Because tRNA genes are usually present in several copies, in some cases even arranged as tandem repeats, they can be considered as a special type of repeated DNA [9], and their evolution can be studied as a repetitive process [66]. A lot of research has been carried out on the mechanisms of replication and propagation of various types of repeated sequences in Prokaryotes [9], such as insertion sequences and transposons, but regrettably it seems that nearly no work has been done to understand how the evolutionary dynamics of repeated DNA can be used to understand the evolution, propagation and maintenance of tRNA copies within genomes.

An important question arises: what underlying mechanisms generate the distribution of bacterial tRNA gene copies seen in figure 5.3? Do tRNA genes present in multiple copies convey a selective advantage to their host or are they just the product of selfish DNA like mechanisms of propagation? Interestingly, as we discussed in chapter 4 (figure 4.2) genome size and total tRNA gene copy number are highly correlated in both Prokaryotes and Eukaryotes, so perhaps the mechanisms that explain genome size evolution might be also taken into account to understand tRNA propagation within genomes. The study of genome size evolution has been very controversial [114], not least because of the iconoclastic ideas behind the selfish DNA hypothesis [18]. This hypothesis assumes that repetitive DNA is maintained and spread throughout genomes due to its inherent replicative properties and does not necessarily confer any selective advantage to its carriers [108]. Under this view, if tRNA genes can be considered to some extent to represent selfish genes, then the tRNA configuration of modern *Escherichia* strains simply reflects a series of selfish propagative events. From a selectionist point of view, it is interesting to notice that bacteria with small genomes tend to present a minimal set of tRNA genes that is non-redundant and that can adequately translate all codons (e.g. [104]). It is plausi-

ble that a small sized bacterial ancestor suffered a series of genome expansions (e.g. through genome duplication, recombination, horizontal gene transfer, etc.) that led to an increased and redundant set of tRNA genes. The fact that many of the tRNA genes analysed are presented as interspersed repeats associated to horizontal transfer events suggests that selfish like mechanisms of propagation are operative. Furthermore, many of the well conserved polycistronic tRNA operons are structured as tandem arrays of repeated tRNA genes where recombination has been observed [9]. Expansions (or reductions) of these repeats probably originated by unequal crossing over, a mechanism that accounts for the generation of large quantities of repetitive DNA. It is possible that the fundamental genomic tRNA backbone is now maintained by selection, but the highly dynamic nature of interspersed tRNAs simply reflects selfish-like processes.

6 Reconstructing ancestral codon sequences

As discussed in the previous chapter, it seems as if the variations in the structure of the genomic tRNA pool of *Escherichia coli* have had no effect on its codon usage. These results are surprising and puzzling. *Escherichia coli* genes can be classified into three groups according to their codon usage [102, 30]. The first group contains highly expressed genes such as ribosomal proteins or elongation factors. These genes show substantial selected codon usage bias. The second group contains most of the genes present in the genome such as housekeeping genes or general metabolism. These genes present moderate selected codon usage. Finally, the third group contains AT rich genes of unknown function and/or that have been horizontally transferred from alien genomes [102, 47, 30]. These gene groups can be placed in an Nc-plot (top figure 6.1) where their relative silent GC content and codon usage can be appreciated. Group three genes are generally assumed to be undergoing amelioration [106], their nucleotide composition changing slowly and approaching that of the current host genome. Groups one and two are under the influence of selection for translational optimisation, with mutation/drift as an antagonistic force. The interesting question is how strong are these forces relative to each other? For example, if selection pressure has increased in the recent evolutionary history of *E. coli*, then groups one and two would be moving downwards in our Nc-plot. Otherwise, if selection has been relaxed, these groups would be actually

moving upwards. There is also the possibility that the system is in equilibrium, selection being finally balanced by mutation/drift, and the gene groups would remain stationary. Exploring these scenarios would give us some clue about how codon usage evolves in *E. coli*, and shed light onto the puzzling issues that have arisen in chapter 5, and onto the more general problem of tRNA and codon evolution in other organisms.

A possible way to explore the codon usage landscape depicted in figure 6.1 would be to identify and align orthologous protein coding sequences from different *E. coli* strains. From the alignment and the estimated phylogenetic tree, ancestral reconstructions could be performed at different nodes in the three, and the ancestral codon bias could be analysed. The reconstructed ancestors could also be placed in the Nc-plot, and the direction of evolution of the different *E. coli* genes in the codon landscape could then be appreciated. This idea is represented schematically at the bottom of figure 6.1. The remainder of this chapter deals with the estimation of ancestral codon sequences for the five *Escherichia-Salmonella* genomes analysed in chapter 5, in the hope of providing a more complete picture of codon evolution.

6.1 Orthologous sequences in the *Escherichia-Salmonella* clade

A total of 2631 orthologs could be identified in the five genomes analysed (figure 6.2). They represent 61% of the genome of *Escherichia coli* K-12 and 59% of *Salmonella typhimurium* LT2. The orthologs are largely colinear across the five genomes, with the main exception of a 318 Kb region at about 1.31 Mb (*Ec*K12 coordinates) that seems to have suffered several inversion events along the *Escherichia-Salmonella* lineages. The 2631 sets of orthologs were aligned and nucleotide substitution rates were estimated by maximum likelihood (table 6.1, PAML package [149]). All genes were assumed to be related under the tree topology obtained in chapter 5

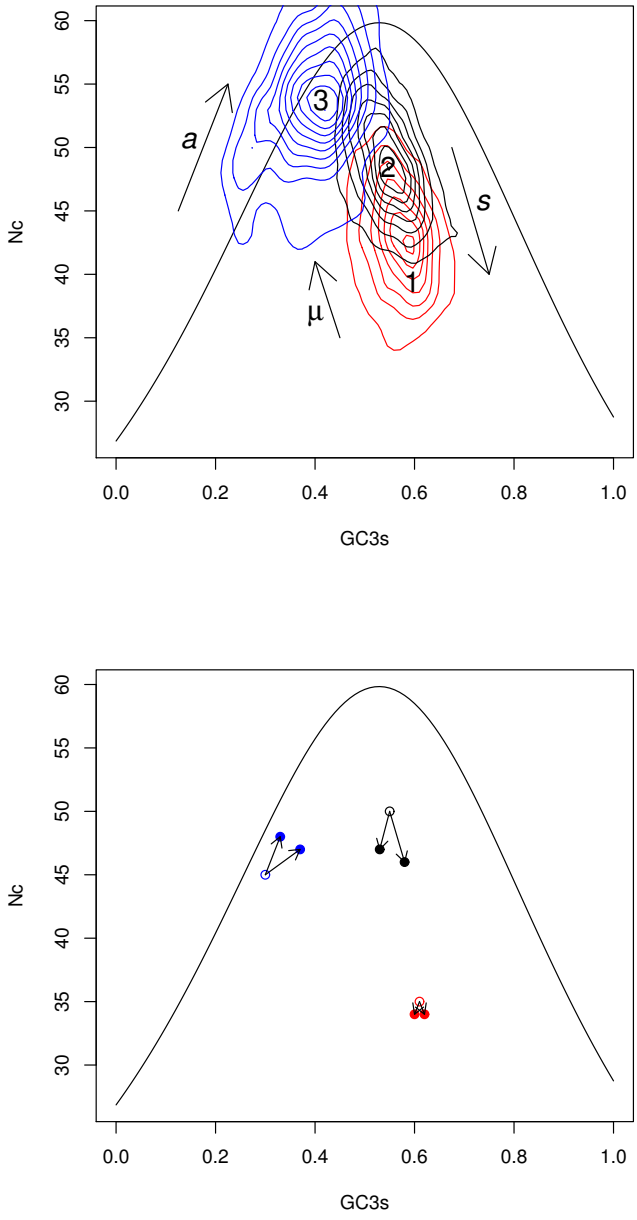

Figure 6.1: Nc-plot of *Escherichia coli* K-12 gene classes.
Top panel: Nc-plot of the three *E. coli* K-12 codon usage groups. The three clouds show the gene population density estimates from a Gaussian bivariate kernel. Red: group 1; black: group 2; and blue: group 3 (redrawn after [30]). The genes classes are assumed to be under the influence of selection (s), amelioration (a) [106] and mutation/drift (μ). Bottom panel: Hypothetical ancestral reconstruction (empty circles) and evolutionary trends for three sets of modern orthologous genes (filled circles).

Table 6.1: Substitution rates for the *Escherichia-Salmonella* clade.

clock?	model[1]	site type	median[2]	quantiles	mean[3]
no	REV + dG=10	1st codon site	0.1421	(0.1011, 0.2162)	0.7937
		2nd codon site	0.1149	(0.08144, 0.1741)	0.6391
		3rd codon site	0.3621	(0.2505, 0.5435)	1.975
yes	REV + dG=10	1st codon site	0.1454	(0.1028, 0.2167)	1.0540
		2nd codon site	0.1213	(0.08568, 0.18170)	0.8768
		3rd codon site	0.4130	(0.2941, 0.6207)	3.029

[1]REV + dG: General time reversible model with rate variation among sites modelled under a discrete gamma distribution with 10 rate categories [147].
[2]All values are x 10^{-8}substitutions per site per Myr.
[3]Mean weighted by sequence length.

(figure 5.2). A discrete gamma model with ten rate categories was used to describe rate variation among sites [147]. The model was fitted with and without assuming a molecular clock. The clock and the no clock model give essentially the same estimates (table 6.1), so the clock model was used for further analysis since it allows the reconstruction of the ancestral root of *Escherichia-Salmonella*. As expected, substitution rates at the third codon positions are higher than at the second and first positions (figure 6.3). Orthologs were grouped according to the clusters of orthologous groups classification [134], and the average Nc and substitution rate per group were calculated (figure 6.4). As expected, genes related to translation and biogenesis (group J) present the lowest Nc values and lowest substitution rates. These genes are highly expressed and present highly optimised codon usage [71, 30]. Some genes cannot be classified into any of the orthologous groups (indicated as '-' in figure 6.4). Most of these genes have been horizontally acquired from foreign genomes [47, 30], and are undergoing amelioration [106]. Accordingly, these genes present the largest Nc values and the higher substitution rates.

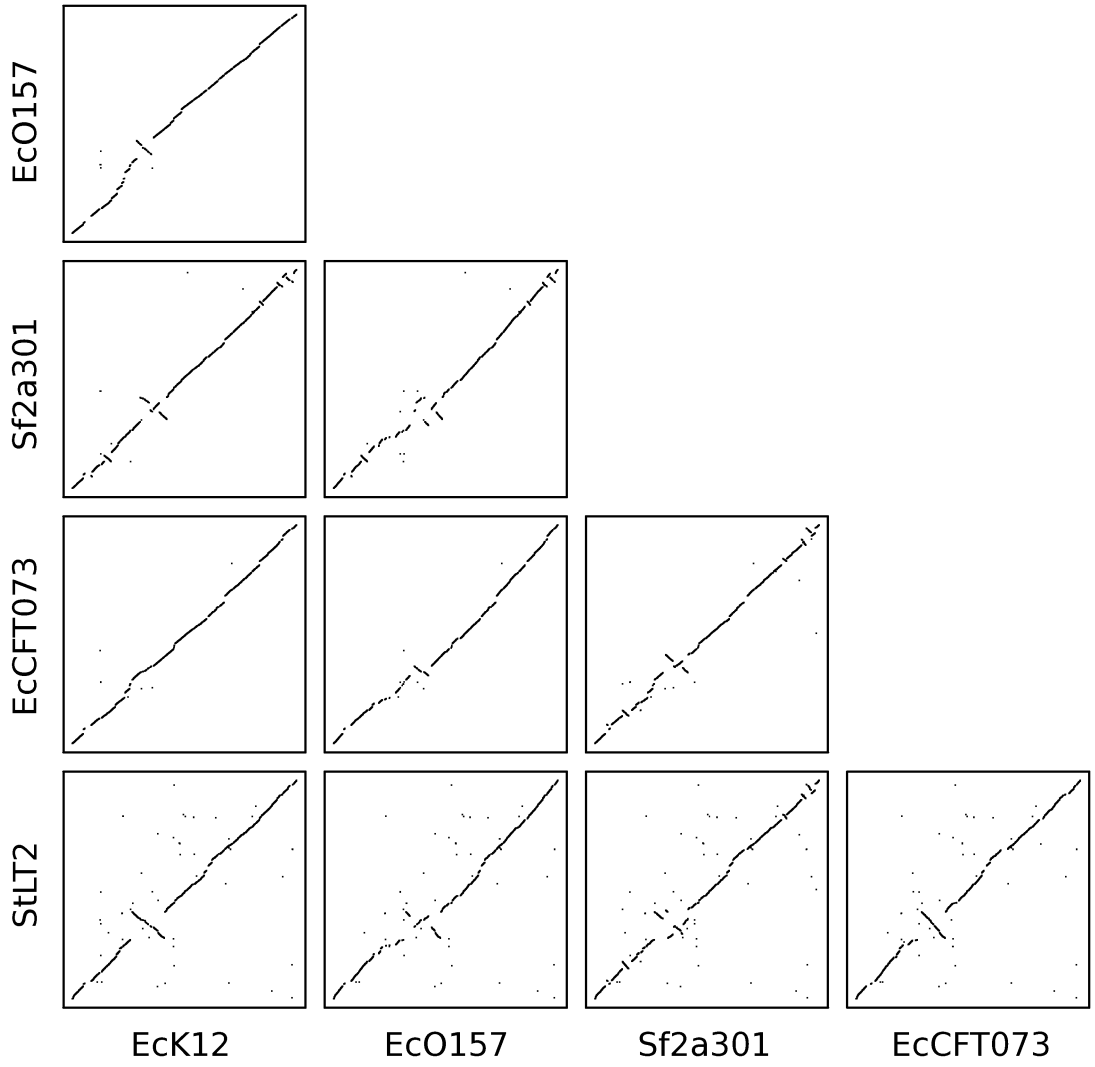

Figure 6.2: Dot plots of orthologous genes in the *Escherichia-Salmonella* clade. Orthologous sequences were identified as the best reciprocal matches from a local alignment search program (BLAT, Jim Kent, unpublished). All translated open reading frames (ORFs) in every genome were cross-searched against all the ORFs in the remaining four genomes. The orthologs are the best reciprocal matches in any two cross BLAT searches. Only sets of orthologs present in all five genomes were kept for the analysis. This technique is shown to produce reliable sets of orthologous sequences [79]. For further analysis, the orthologous protein sequences were aligned using MUSCLE [37, 36], and codon alignments were obtained using PAL2NAL [133] using the MUSCLE alignment as a reference.

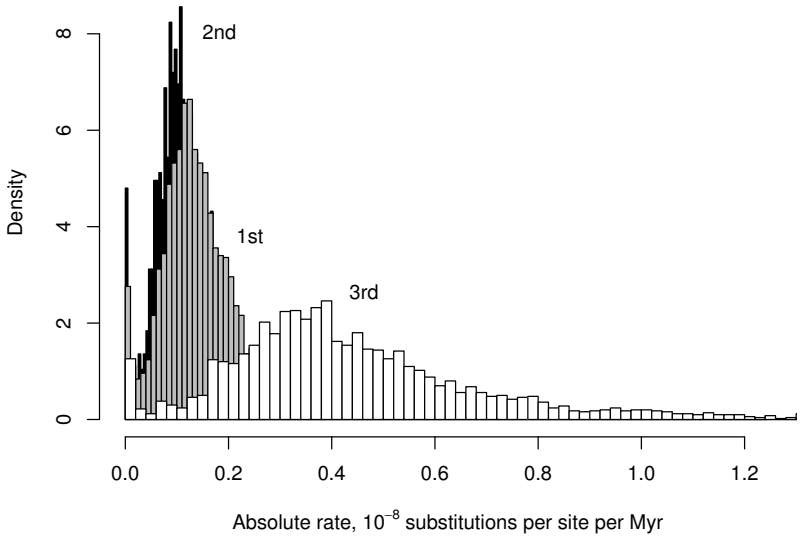

Figure 6.3: Absolute nucleotide substitution rates in the *Escherichia-Salmonella* clade.

A total of 2631 orthologous genes were analysed. Grey: 1st; black: 2nd; and white: 3rd codon positions. When a discrete rate variation model is estimated (such as the discrete gamma), each site is assigned to a particular rate class. The value of each rate class is relative, such that the average of all rate classes (10 in this case) is one [147, 151]. The relative rate values have no units, but they can be scaled using a molecular clock to obtain the absolute rates. Including rate variation among sites is fundamental to obtain reliable estimates of substitution rates and branch lengths in phylogenetic trees [148].

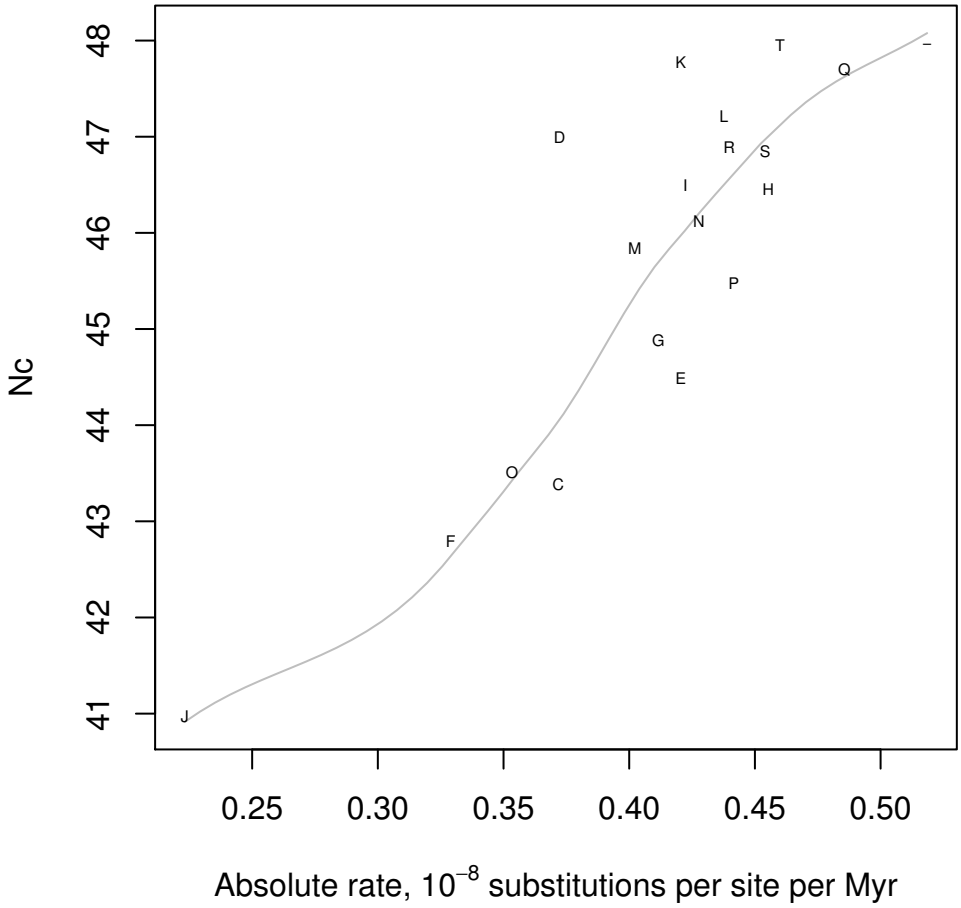

Figure 6.4: Codon usage vs average substitution rates among protein functional groups.

Orthologs were grouped according to the functional classification of the clusters of orthologous groups classification [134]. Codes are: (J) translation, ribosome structure and biogenesis; (K) transcription; (L) DNA replication, recombination and repair; (D) cell division and chromosome partitioning; (O) post-translational modification, protein turnover, chaperons; (M) cell envelope biogenesis, outer membrane; (N) cell motility and secretion; (P) inorganic ion transport and metabolism; (T) signal transduction mechanisms; (C) energy production and conversion; (G) carbohydrate transport and metabolism; (E) amino acid transport and metabolism; (F) nucleotide transport and metabolism; (H) coenzyme metabolism; (I) lipid metabolism; (Q) secondary metabolites biosynthesis, transport and catabolism; (R) general function prediction only; (S) function unknown; (-), not in COGs. Line is a running mean smoother.

6.2 Ancestral codon sequences

The ancestral orthologous sequences at every node in the tree were inferred as described by Yang *et al*. [152]. In this method, evolutionary parameters (such as branch lengths) are estimated by maximum likelihood, and then an empirical Bayes approach is used to estimate the posterior probabilities for each character at each site in a sequence in a given node in the three. The character with the highest posterior is chosen as the estimated ancestral character. The ancestor estimated in this manner is called the maximum likelihood (ML) reconstruction. Maximum parsimony reconstruction of ancestral characters is also possible, but they usually perform worse than the ML estimates [152, 26], so this method will not be considered here. Figure 6.5 shows the ML reconstruction of the root of the *Escherichia-Salmonella* tree in an Nc-plot, and the direction of codon evolution towards the *Ec*K12 genome. The orthologs were assumed to follow the tree estimated previously with nucleotide substitutions following a reversible model with discrete gamma variation among sites (this is model REV + dG=10 in table 6.1). The ML reconstructions seem to suggest that the last common ancestor of the *Escherichia-Salmonella* clade presented substantially more biased codon usage and nucleotide composition. However, this result should be taken with scepticism. Although the ML ancestral reconstruction is usually the most accurate, it is nearly guaranteed to be biased [142]. Intuitively, it is easy to see why this should be so. Let us imagine a sequence composed uniquely of characters of type X and Y. Let assume the extreme case where all sites have a posterior probability of 0.6 for character X in the reconstruction. Then the reconstructed ancestor will be composed exclusively of X characters, when it is easy to see that it should actually be composed of 60% X's and 40% Y's.

Sometimes we are more interested in certain properties of the ancestral sequence (such as GC content) rather than on the exact character configuration of the ancestral sequence itself. When this is true, it is usually best to calculate the posterior probabilities of each one of the possible reconstructions (this is simply the prod-

uct of the posteriors at individual sites, assuming sites evolve independently), and then the parameter of interest is calculated for each possible reconstruction. The weighted mean of the parameter by the posterior probability of each reconstruction is then a good estimator of the value of the ancestral parameter. The problem with this approach is that it is computationally unfeasible but for the smallest sequences. A more feasible approach is to simulate a representative set of ancestral sequences from the posterior probabilities at each site, then the parameter of interest is estimated for each simulated sequence, and the ancestral value of the parameter is simply the arithmetic mean of the simulated values. This randomised reconstruction based on the posteriors has been coined *Bayesian reconstruction* by Williams *et al*. [142]. Figure 6.6 shows the Bayesian codon reconstruction at the root of the *Escherichia-Salmonella* clade in an Nc-plot. The results are substantially different from the ML case. It seems that the last common ancestor of the *Escherichia-Salmonella* clade had substantially less biased codon use, and slightly richer GC composition. These results seem more coherent than the ML case.

The evolutionary paths shown in figure 6.6 are over simplified since they only show two actual points for each ortholog. The orthologous sequences might not have been evolving in such straight lines. Furthermore, we are also interested in observing the most recent codon evolution in the *Escherichia* clade (so we could understand the recent miss-shift in codon - tRNA coadaptation observed, chapter 5). Figure 6.6 could be redrawn to show all the modern sequences and the reconstructions at each one of the four ancestral nodes in tree. However, this would result in an unreadable graph. A simple approach is to calculate average Nc and GC3s values according to protein functional groups, and plot the average parameters instead for each node in the tree in the Nc-plot. The result of this exercise is shown in figure 6.7. Several important features about the evolution of codon usage in the five genomes analysed can be learnt. First, the last common ancestor of *Escherichia-Salmonella* seemed to have had an average GC content that was intermediate to

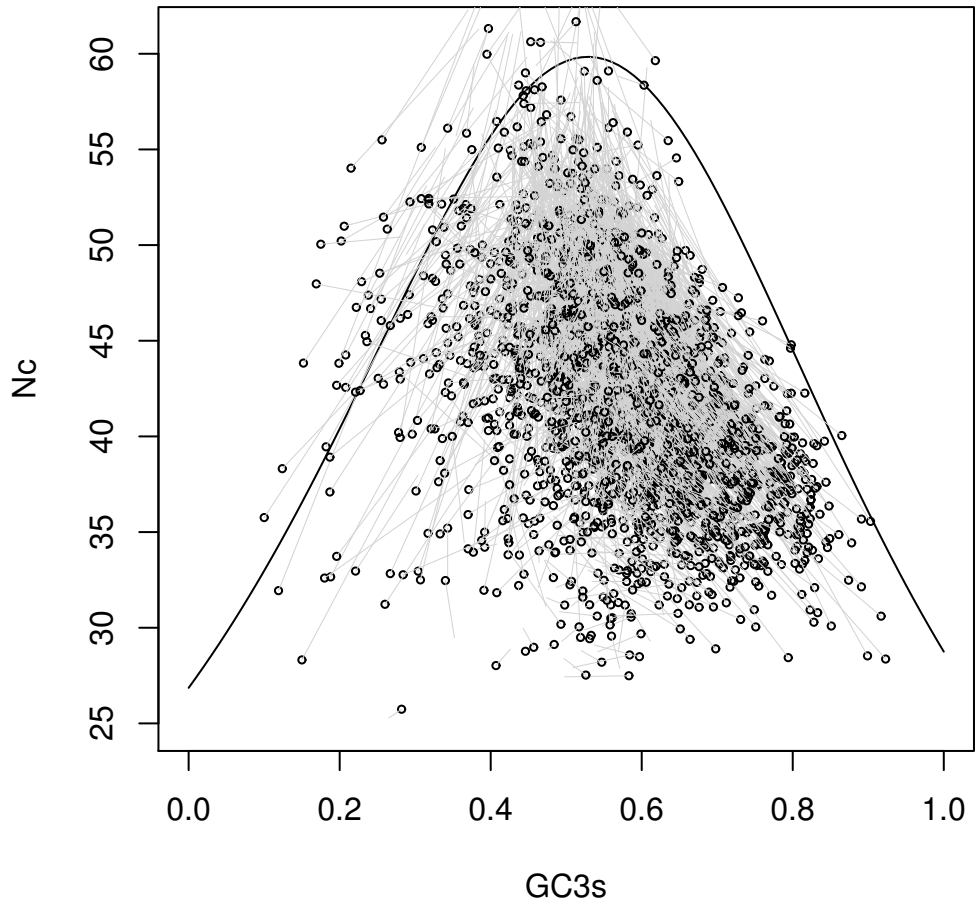

Figure 6.5: Maximum likelihood ancestral codon reconstruction.
The empty circles show the location of each reconstructed *Escherichia-Salmonella* orthologous ancestral sequence. The lines show the estimated direction of evolution towards the modern *Ec*K12 genome (the line tips). Most of the lines point upwards, which would suggest that the *Escherichia-Salmonella* ancestor had a more codon biased genome.

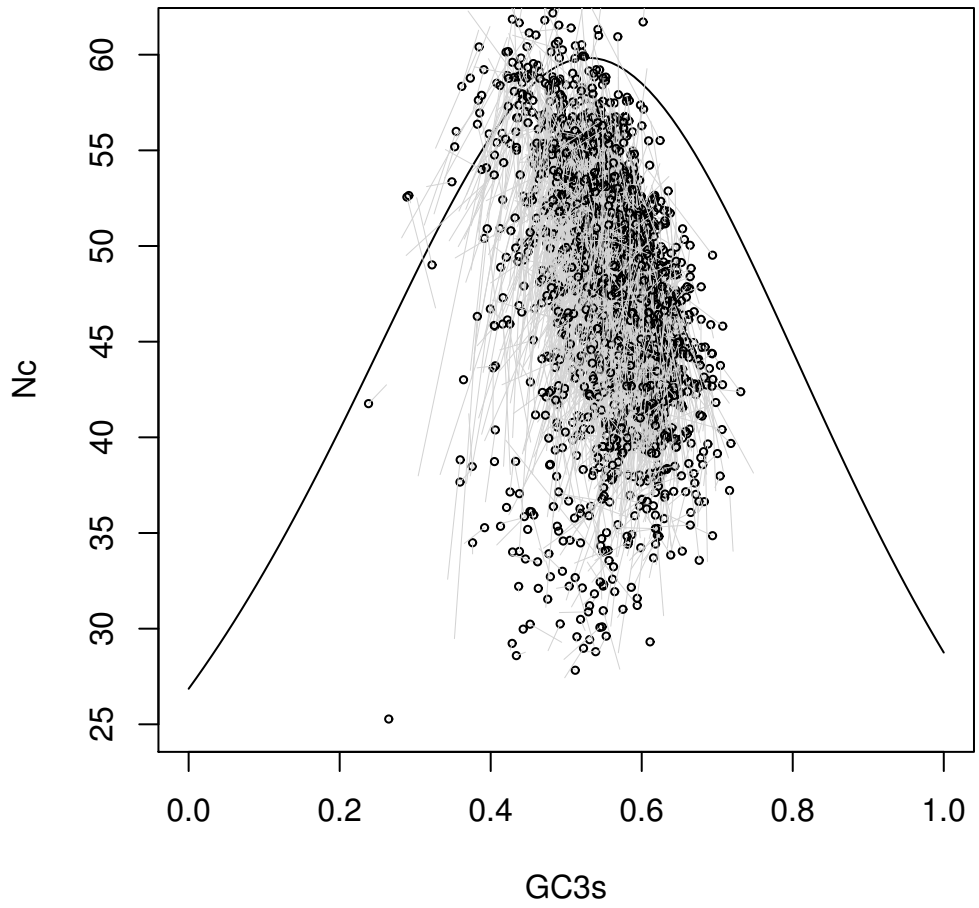

Figure 6.6: Bayesian ancestral codon reconstruction.
The empty circles show the location of each reconstructed *Escherichia-Salmonella* orthologous ancestral sequence. The lines show the estimated direction of evolution towards the modern *Ec*K12 genome (the line tips). The evolutionary parameters were estimated by maximum likelihood and ten ancestral sequences were obtained by sampling randomly from the posterior probabilities of each character at every site in the root node. Nc and GC3s values were computed from the randomised sequences and the values were averaged. Most of the lines point downwards, which would suggest that the *Escherichia-Salmonella* ancestor had a less codon biased genome. Contrast this with figure 6.5.

that of the modern *Escherichia* and *Salmonella* genomes. This suggests that during the past 100 Myr of evolution, *St*LT2 has become slightly GC richer, while *Escherichia* spp. have become AT richer. However, whether this is a real fact, or an artifact of the reconstruction method employed cannot be addressed with current data. Further analyses, perhaps including more divergent genomes and using more complex models (i.e. non-stationary and non-homogeneous e.g. [154]) are needed to solve this issue. It can also be seen that since their divergence about 100 Myr ago, both *Escherichia* and *Salmonella* have undergone codon usage optimisation. However, it is striking to note that in the most recent evolutionary history of *Escherichia coli*, a sub-optimisation of codon usage seems to have taken place. This sub-optimisation seems to be less marked in *Ec*K12 (not shown) than in the other *Escherichia* genomes analysed. In a sense, these results seem to agree with our previous findings that codon usage and the genomic tRNA pool have been decoupled in the recent evolution of *Escherichia* spp. (table 5.5, chapter 5). However, the reasons for this sub-optimisation are not clear at the moment. Several speculative explanations can be put forward: a recent reduction in effective population size in *Escherichia* spp. with the subsequent segregation of slightly deleterious mutants [122]; the dynamic nature of the genomic tRNA set has created shifts in selective pressure and codon usage is drifting to a new point of equilibrium; or there might be problems with the ancestral reconstruction. More research is needed to dig deeper into this issue. It is hoped these findings will entice other researchers to follow up this interesting area of research.

Williams *et al* [142] performed extensive tests comparing ancestral reconstruction of simulated sets of proteins under parsimony, maximum likelihood and Bayesian methods. They found that the ML method had the highest accuracy, but it was consistently biased in estimating parameters such as the thermostability, etc. The Bayesian method, although slightly less accurate, did not suffer from any apparent bias. Their work was focused on simulated amino acid sequences, so it would be

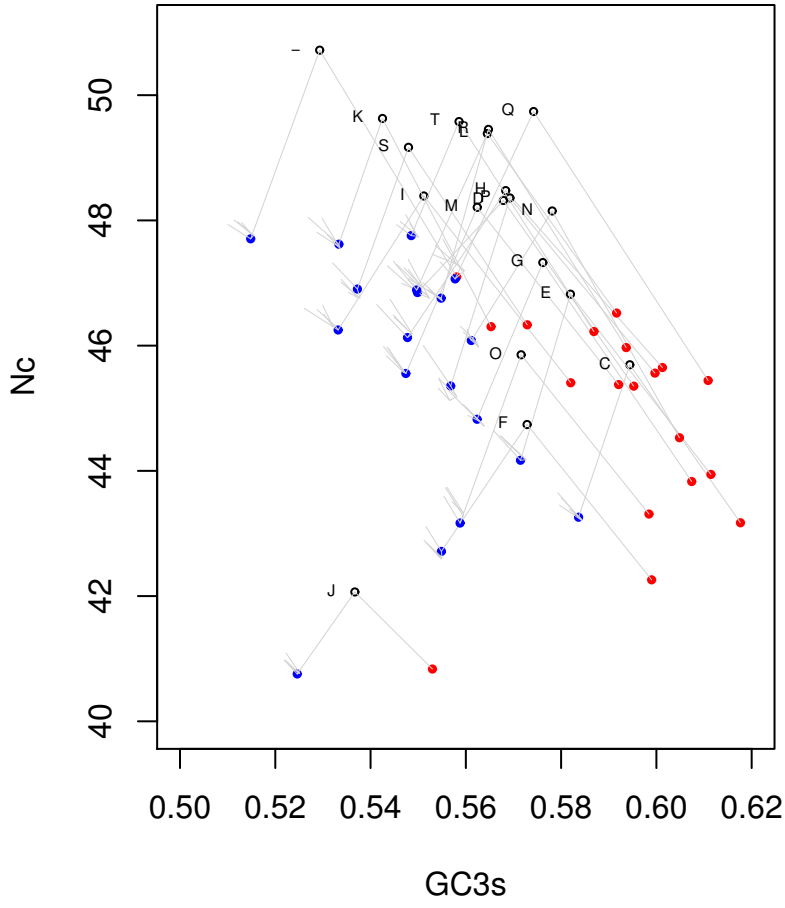

Figure 6.7: Bayesian ancestral codon reconstruction for the *Escherichia-Salmonella* clade.

Nc and GC3s values were averaged across clusters of orthologous groups for each node in the tree. The empty circles show the location of the reconstructed *Escherichia-Salmonella* ancestor (the root of the tree). The red dots show the location of the modern *St*LT2 genes. The blue dots show the location of the reconstructed ancestor for the four *Escherichia* genomes analysed. Because the ancestral reconstruction assumes homogeneity and stationarity [152], the ancestor is forcibly estimated as having an intermediate nucleotide composition compared to the modern strains. The modern *Escherichia* genes for each genome are located at the tips of the grey lines projecting away from the blue dots. The seemingly small sub-optimisation in modern *Escherichia* strains could be due to the segregation of slightly deleterious mutants [122]. Ten sequences were simulated for each ortholog at every one of the four internal nodes in the three. The simulations were performed as in figure 6.6.

interesting to perform a similar analysis on nucleotide sequences. We are currently performing simulations of nucleotide sequences and assessing the performance of the Bayesian method.

Finally, it must be stressed here again what the term Bayesian reconstruction means. In the ML reconstruction the evolutionary parameters are estimated by maximum likelihood, and then the characters at every site are chosen as to maximise the posterior probability of the reconstructed sequence. A similar technique would be to estimate the evolutionary parameters by a fully Bayesian approach using Markov Chain Monte Carlo integration (for example, as implemented in Mr. Bayes [59, 58, 121]), but again, the characters at every site would be chosen as to maximise the posterior of the ancestor. Although I am not aware of any works that have tested the performance of ancestral reconstruction from the MCMC Bayesian approach, I would expect it to suffer from the same drawbacks as the ML reconstruction. Hence, here, Bayesian reconstruction does not refer to the process of estimation of the evolutionary parameters but rather to the probabilistic nature of reconstructed ancestral properties from the posterior averages. It is also important to note here that the ancestral reconstruction at the most ancestral node in the tree is not completely reliable. The level of uncertainty is substantial. The branch leading from the root of the tree to the most internal *Escherichia* node is considerably long. Because a few sites seem to be saturated, part of the phylogenetic signal has been lost. The current sequencing projects in *Escherichia* and *Salmonella* will provide new data that could be used to improve this analysis. The addition to the tree of more *Salmonella* species, and if possible, of *Escherichia* strains with more ancient divergence times, could improve the ancestral reconstruction at the root node.

7 Codon usage and genome evolution

The neutral theory of molecular evolution maintains that most evolution at the molecular level is shaped by mutation and the effects of random genetic drift, without the intervention of natural selection [76]. Random drift adds a factor of uncertainty to evolution. The fate of a mutant in a population depends not only on its effect on the organism fitness, but also on the inherent randomness of allele propagation throughout generations. As we discussed in chapter 2, the probability of fixation of a novel mutant in a population depends on the effective population size. If the population number is sufficiently small, a slightly advantageous mutant alleles will spread in the population in the same manner as a neutral mutant. This has important implications for the evolution of codon usage. Since the selection coefficients acting on codon usage are small [54], selected codon bias should only be apparent in organisms with sufficiently large populations. Figure 7.1 shows the equilibrium frequency of an optimal codon as a function of effective population size under weak selection. The role of effective population size on the evolution of codon usage has been well discussed in the literature [127, 17]. However, what is surprising, is that it seems no workers have actually tested this idea empirically. Perhaps with the exception of a paper by Akashi [4], no actual data seems to have been produced comparing actual population sizes with actual estimates of selection on codon usage. This is, in my opinion, one of the largest gaps remaining in codon

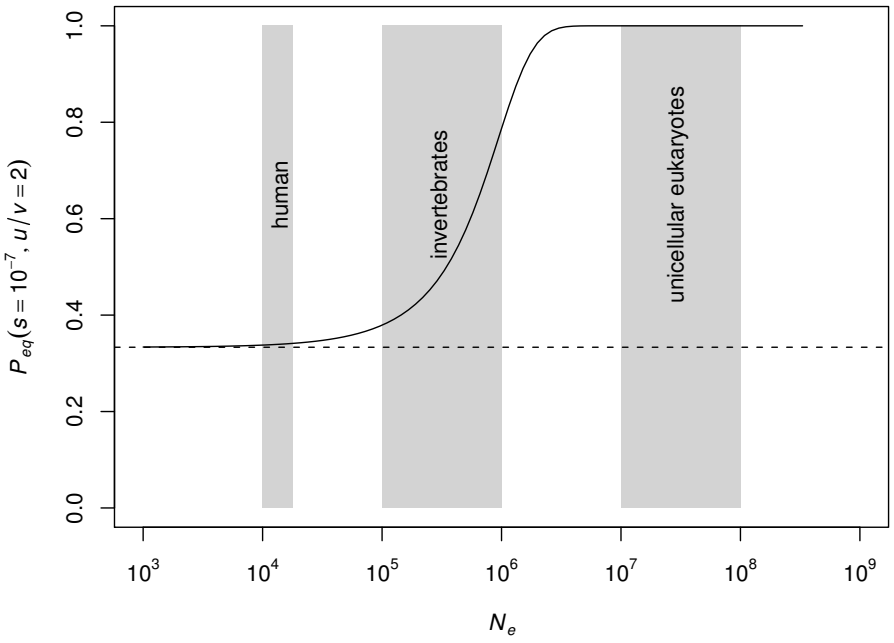

Figure 7.1: Equilibrium frequency of an optimal codon vs effective population size in Eukaryotes.
The curve represented in the figure is from rearranging equation 2.4, chapter 2. The grey rectangles show estimated effective population number intervals for some eukaryotic groups [129, 52, 92]. Estimates of selection on codon usage are in the order of 10^{-6} to 10^{-9} (chapter 2, [54, 97]), the value of $s = 10^{-7}$ is shown for illustrative purposes only. Mutational bias (u/v) for the substitution of C with T is about 1.41 to 1.95 in yeast (table 2.1). The dashed line shows the equilibrium frequency of the codon under no selection.

usage research.

7.1 Codon usage and population size in Eukaryotes

The following is an analysis of selection on codon usage and effective population size estimates in Eukaryotic genomes. This analysis is rather exploratory in nature, and it will be presented here for illustrative purposes only, since some of the assumptions of the analysis are clearly violated [29]. It will serve to summarise the

ideas discussed in this work, and hopefully, it will help clear a path towards future research in this area.

Random drift causes the non preferential fixation of certain alleles in a population and the elimination of others. Because the time it takes for an allele to become fixed is proportional to the population size, new mutants at individual sites tend to become fixed relatively quickly in small populations, while individual sites tend to remain polymorphic in larger populations [76, 122]. Thus, random drift causes a reduction of genetic diversity, or polymorphism, in small populations [76]. This provides a way to use molecular data to estimate population size. Using polymorphism levels from gene alignments, it is possible to estimate the quantity $N_e u$, which is the product between the effective population size (N_e) and the mutation rate per generation (u). Because mutation rates vary only by around one to two orders of magnitude from Prokaryotes to Eukaryotes [92, 91], this quantity provides a reasonable measure of N_e, which in turn varies by several orders of magnitude. Recently, Lynch and Conery [92] used this idea to obtain estimates of $N_e u$ for several Eukaryotic and Prokaryotic genomes. This provides us with a unique opportunity to compare the $N_e u$ values with estimates of selection on codon usage. This will tell us whether, as should be expected, organisms with small population sizes show signs of reduced selected codon usage bias. Lynch and Conery only made available their $N_e u$ data for Eukaryotic genomes, so the following analysis will be focused solely on this group.

As we saw in chapter 2, the common Baker's yeast is a model organism in codon usage research. Good quality microarray data is available that allowed us to classify the yeast genes into expression bins, and then calculate the intensity of selected codon bias in these binned genes (chapter 2). These bins can be regarded as expression categories, and the genes contained within them can be used as a reference to find the corresponding orthologs in other eukaryotic genomes. If the orthologous genes are assumed to have conserved relative expression levels across species, then we can apply the technique described in chapter 2 to estimate translational selection

in any genome where a sufficiently large number of orthologous yeast genes can be identified. As it turns out, performing reciprocal searches of orthologs between genomes as disparate as yeast and *Arabidopsis thaliana* can yield enough data to perform the analysis. Figure 7.2 shows the results of this exercise. The 77 expression bins obtained for the yeast genome (figure 2.7, chapter 2) were collapsed into 11 expression categories, sorted according to increasing expression levels, in order to calculate the $\hat{S}$ values. Since $\hat{S}$ is a function of two binomially distributed variables, the frequency of optimal codons in highly and lowly expressed genes, its variance is inversely proportional to the number of codons analysed, so relatively large bins containing thousands of codons are desirable. As expected, $\hat{S}$ is positively correlated with increasing expression category. What is surprising though, is that the trend is very similar for most species. Furthermore, orthologs belonging to the highest expression category also show the larger $\hat{S}$ values across *all* genomes analysed. It must be stressed here again that this analysis assumes that expression levels are conserved across all species. Also it is assumed that mutation rates between highly and lowly expressed genes are about the same (u/v equation 2.4) so a meaningful estimate of S can be obtained. This is why this analysis should be regarded with a bit of scepticism. However the results shown in figure 7.2 are very suggestive, and a more detailed analysis should confirm them.

The technique used here to estimate S is symmetrical for codons belonging to synonymous family two. This means that if a codon c_1 has a selective value S_1 then the selective value against the complementary codon c_2 is simply $S_2 = -S_1$. This property allows the estimation of S irrespective of whether we know beforehand which codon is the optimal one. In fact, it allows the identification of the optimal codon by computing S itself. If a negative value is obtained, then the optimal codon should be the complementary one. The only exception to this rule is when S is zero. This can be determined by bootstrapping as shown in chapter 2 for the yeast example. Table 7.1 shows estimated S values for all nine optimal codons belonging to

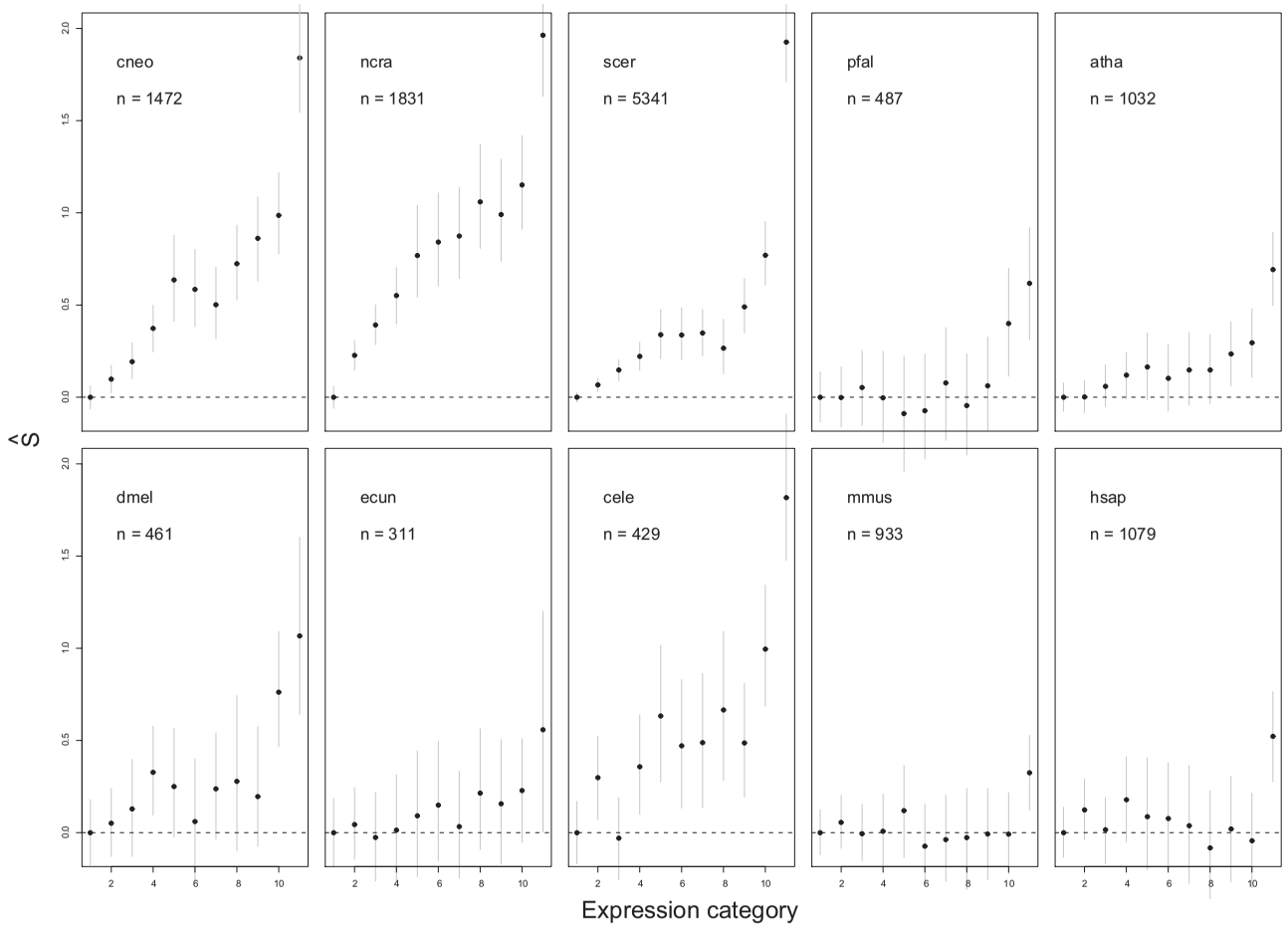

Figure 7.2: Estimated *S* values across expression categories for several Eukaryotic genomes. Organism codes are: mmus, *Mus musculus*; hsap, *Homo sapiens*; pfal, *Plasmodium falciparum*; atha, *Arabidopsis thaliana*; cele, *Caenorhabditis elegans*; ecun, *Encephalitozoon cuniculi*; dmel, *Drosophila melanogaster*; ncra, *Neurospora crassa*; scer, *Saccharomyces cerevisiae*; cneo, *Cryptococcus neoformans*. The values on the upper left corner of each panel are the number of orthologs analysed. Orthologs were identified as described in chapter 6 for the *Escherichia-Salmonella* clade. The vertical bars are the 95% CI from 1,000 bootstrap replicates.

the synonymous family two. The three most highly expressed bins from the original yeast analysis (figure 2.7) were collapsed into one putatively highly expressed orthologous category to obtained the values shown in table 7.1. It is interesting to note that optimal codons are not always the same across the genomes analysed. For example codon CAA coding for Gln is preferred in *S. cerevisiae*, while its counterpart CAG seems to be consistently preferred by the other genomes. Estimates of S are statistically different from zero for most amino acids in most genomes. For those cases where $\hat{S}$ is indistinguishable from zero, the optimal codon was arbitrarily chosen as the one with a positive $\hat{S}$ value. This does not seem to cause important biases in the estimated average of $\hat{S}$ presented in figures 7.2 and 7.3.

It is time to draw attention to the problem of effective population size. Figure 7.3 shows the estimated S values for the most putatively highly expressed orthologs in the ten Eukaryotic genomes analysed (these are equivalent to the three most highly expressed bins in the yeast analysis, chapter 2), plotted against Lynch and Conery $N_e u$ values. The large mammalian genomes (*Mus musculus* and *Homo sapiens*) show low $\hat{S}$ values, while the fast growing, fungal genomes (*Neurospora crassa*, *Saccharomyces cerevisiae* and *Cryptococcus neoformans*) show the largest values. The two multicellular invertebrates (*Caenorhabditis elegans* and *Drosophila melanogaster*) and the plant genome (*Arabidopsis thaliana*) show large variation in $\hat{S}$ values. The two intracellular parasites (*Encephalitozoon cuniculi* and *Plasmodium falciparum*) show relatively low values. All genomes analysed seem to show $\hat{S}$ values that are statistically different from zero, as suggested by the non-parametric bootstrap intervals. The presence of selected codon bias in humans (and mammals in general) has been controversial, with reports in favour and against the presence of translational selection (see for example [41, 136]). Recent work by Urrutia and Hurst [137] have shown that there is evidence of selected codon usage in humans. The results presented here suggest that weak selection is operative in the most putatively highly expressed genes in mammals.

Table 7.1: Estimated S values for several Eukaryotes

		Cryptococcus neoformans				*Neurospora crassa*				*Saccharomyces cerevisiae*				*Plasmodium falciparum*				*Arabidopsis thaliana*			
Aa	**Codon**	**tRNA**	$\hat{S}$	**2.5%**	**97.5%**	**tRNA**	$\hat{S}$	**2.5%**	**97.5%**	**tRNA**	$\hat{S}$	**2.5%**	**97.5%**	**tRNA**	$\hat{S}$	**2.5%**	**97.5%**	**tRNA**	$\hat{S}$	**2.5%**	**97.5%**
Phe	TTT	0				0				0				0				0			
	TTC	5	1.58	1.19	2.12	12	2.41	1.84	3.12	10	2.21	1.99	2.45	1	0.53	0.12	0.92	16	0.68	0.34	1.08
Tyr	TAT	0				0				0				0				0			
	TAC	4	2.54	1.78	4.11	11	1.73	1.33	2.29	8	2.79	2.36	3.26	1	1.23	0.77	1.60	76	1.06	0.63	1.57
Cys	TGT	0				0				0	1.22	0.53	2.29	0				0	0.05	-0.35	0.48
	TGC	3	0.26	-0.22	0.76	7	1.92	1.19	3.53	4				0	1.10	0.52	1.61	15			
His	CAT	0				0				0				0				0			
	CAC	4	2.27	1.59	3.17	9	2.05	1.60	2.78	7	1.98	1.68	2.32	2	1.03	0.42	1.58	10	0.87	0.43	1.40
Gln	CAA	2				3				9	4.16	3.50	5.67	1	0.42	-0.10	1.12	8			
	CAG	3	1.54	1.18	1.95	11	2.89	2.38	3.64	1				1				9	0.50	0.16	0.93
Asn	AAT	4				10				10				1				16			
	AAC	1	3.11	2.50	4.26	2	3.29	2.67	4.80	7	2.83	2.53	3.16	2	1.35	0.98	1.69	13	0.88	0.64	1.14
Lys	AAA	7				24				14				2				18			
	AAG	0	2.98	2.58	3.62	0	3.44	2.64	4.96	0	2.35	2.17	2.53	0	0.38	0.16	0.59	0	1.06	0.79	1.38
Asp	GAT	1				17				16				1				23			
	GAC	3	2.13	1.78	2.62	5	0.38	0.08	0.73	14	1.10	0.92	1.27	1	0.29	-0.13	0.62	12	0.34	0.03	0.63
Glu	GAA	9				23				2	3.42	3.02	3.93	1	0.60	0.19	1.15	13			
	GAG	0	1.73	1.48	2.10	0	2.67	2.23	3.28	0				0				1	0.56	0.23	0.93

		Drosophila melanogaster				*Encephalitozoon cuniculi*				*Caenorhabditis elegans*				*Mus musculus*				*Homo sapiens*			
Aa	**Codon**	**tRNA**	$\hat{S}$	**2.5%**	**97.5%**	**tRNA**	$\hat{S}$	**2.5%**	**97.5%**	**tRNA**	$\hat{S}$	**2.5%**	**97.5%**	**tRNA**	$\hat{S}$	**2.5%**	**97.5%**	**tRNA**	$\hat{S}$	**2.5%**	**97.5%**
Phe	TTT	0				0	0.06	-0.37	0.64	0				0				0			
	TTC	8	1.18	0.47	1.97	1				16	2.93	2.30	4.37	7	0.28	0.00	0.56	14	0.55	0.25	0.88
Tyr	TAT	0				0	0.03	-0.35	0.44	0				0				1			
	TAC	9	1.30	0.74	2.04	1				19	2.06	1.51	2.70	11	0.26	-0.10	0.71	11	0.21	-0.08	0.55
Cys	TGT	0				0	0.46	-0.03	0.91	0				0	0.04	-0.39	0.49	0			
	TGC	7	1.28	0.30	3.42	1				13	1.77	1.20	2.53	56				30	0.16	-0.25	0.63
His	CAT	0				0				0				0				0			
	CAC	5	0.58	0.03	1.35	1	0.19	-0.26	0.66	17	1.45	1.06	1.87	10	0.20	-0.21	0.62	12	0.85	0.45	1.26
Gln	CAA	4				1				18				7				11			
	CAG	8	1.05	0.34	2.09	1	0.75	-0.04	1.87	7	0.05	-0.29	0.37	10	0.03	-0.40	0.50	21	0.52	0.13	1.01
Asn	AAT	9				2				20				15				33			
	AAC	7	1.59	1.03	2.10	1	0.48	-0.08	0.91	16	2.65	2.14	3.22	12	0.22	-0.12	0.56	16	0.73	0.39	1.09
Lys	AAA	13				1				33				30				22			
	AAG	0	1.47	0.76	2.39	0	1.01	0.39	1.74	0	2.72	2.16	3.35	0	0.31	0.01	0.60	0	0.61	0.32	0.93
Asp	GAT	11				1				22				16	0.27	0.02	0.54	10			
	GAC	5	0.20	-0.12	0.54	1	0.17	-0.27	0.69	17	1.08	0.79	1.40	8				14	0.16	-0.21	0.56
Glu	GAA	12				1				20				11				8			
	GAG	0	0.79	0.16	1.38	0	0.68	-0.27	1.62	0	1.48	1.15	1.80	0	0.14	-0.20	0.47	0	0.48	0.13	0.87

tRNA, number of tRNA genes that recognise the given codon; 2.5% and 97.5% are the percentiles from 1000 bootstrap samples on $\hat{S}$.

Finally, $\hat{S}$ is simply the log odds ratio of optimal codons in highly expressed vs lowly expressed genes. If the differences are due to variation in mutation rates between both groups, then spurious (and statistically significant) $\hat{S}$ values might be obtained (see the example in figure 4.8 on page 75). Some comfort could be gained if the direction of optimal codon preferences would correlate with the genomic tRNA pool. In this sense, the S_t test becomes useful. Figure 7.4 shows the estimated S_t values from chapter 4 (table 4.1) plotted against the $\hat{S}$ values of this chapter for the ten genomes being considered. It can be seen that both sets of values are positively correlated ($R = 0.91$) giving strong support to the notion that the $\hat{S}$ values truly reflect the effects of natural selection on codon usage.

7.2 Genome complexity and codon usage evolution

Lynch and Conery [92] have proposed that while Prokaryotes evolved towards multicellular Eukaryotes, the consequent increase in organism size was linked to a dramatic reduction in population size. Because random genetic drift is stronger in smaller populations, this allowed the accumulation of genomic features that would have been eliminated by the action of natural selection in larger populations. These features include the evolution and expansion of introns, gene duplication, and the accumulation of repetitive DNA, mechanisms that account for genome size expansion. Lynch and Conery present convincing evidence showing that indeed population size and genome size are inversely correlated from Prokaryotes to Eukaryotes. These findings suggest that any form of weak selection tends to be overridden by random drift in organisms with large genomes, and this must include selection on codon usage. Lynch and Conery ideas are very appealing, and could provide a nice framework onto which the action of translational selection could be understood.

The findings presented in this work permit the following conjecture of how codon

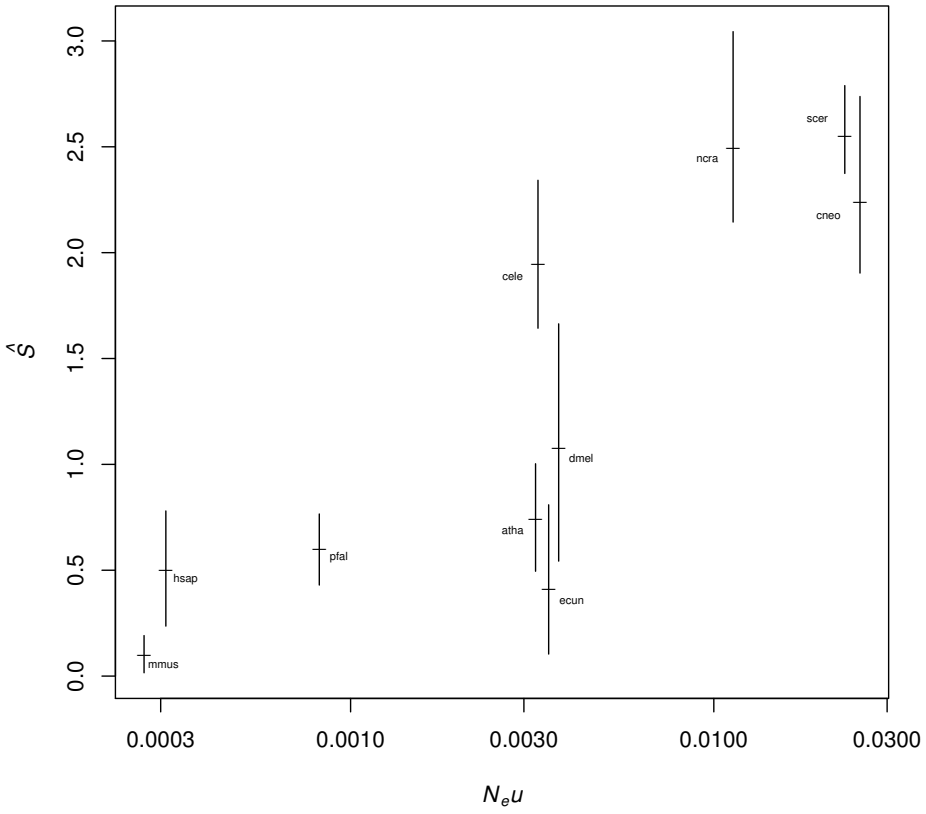

Figure 7.3: Estimated S values for several Eukaryotic genomes vs $N_e u$. Organism codes as in figure 124. The horizontal bars are the mean values, and the vertical bars are the 95% non-parametric bootstrap intervals. The three most highly expressed bins from the 77 original bins in the yeast analysis (figure 2.7) were collapsed into a putatively highly expressed orthologous category.

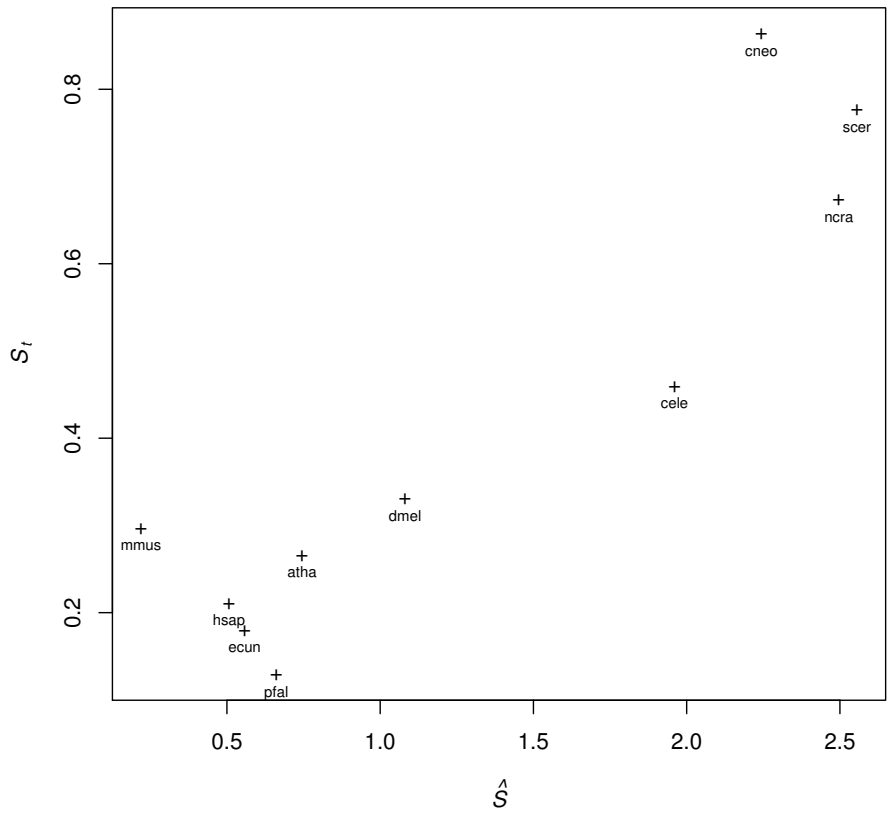

Figure 7.4: S_t vs $\hat{S}$ values for several Eukaryotic genomes.
Organism codes as in figure 7.2. S_t values from table 4.1 (page 81 on chapter 4).

usage optimisation might have evolved. First a hypothetical ancestor (figure 7.5) with a small genome and a reduced set of tRNA genes suffered a series of genome expansions that led to an increased set of tRNA genes. As successive expansions took place, the redundancy of the tRNA set increased and selective pressure for codon optimisation started to be operative. In some instances, selection for fast growth rate might have caused an increase in genomic tRNA content, without an increase in genome size. From this, the first medium genome sized bacterial genomes originated, similar to *E. coli*. Further expansions might have produced the first eukaryotic genomes such as yeast, where codon optimisation is highly developed. As genome size increased further, the concomitant reduction in population size hindered the action of selection on codon usage, generating the large modern genomes such as those of mammals. Selection for reduction of genome size or tRNA redundancy in certain non-free living organisms would invert the process. This conjectural model might be used as a plausible framework onto which research into codon usage may be devised. How and why organisms would start to move about this genomic landscape, and when they should remain stationary are interesting questions that need to be addressed. For example, bacterial genomes are ancient compared to the mammalian counterparts. Why bacteria have remained stationary in the genomic landscape is probably related to the successful colonisation of the ecological niches they occupy. Changes in the lifestyle of organisms might entice their movement about the landscape. It is assumed that some higher Metazoans started to colonise new ecological niches that were associated to larger organism size, this was associated to a reduction in population size and hence an increase in genome complexity.

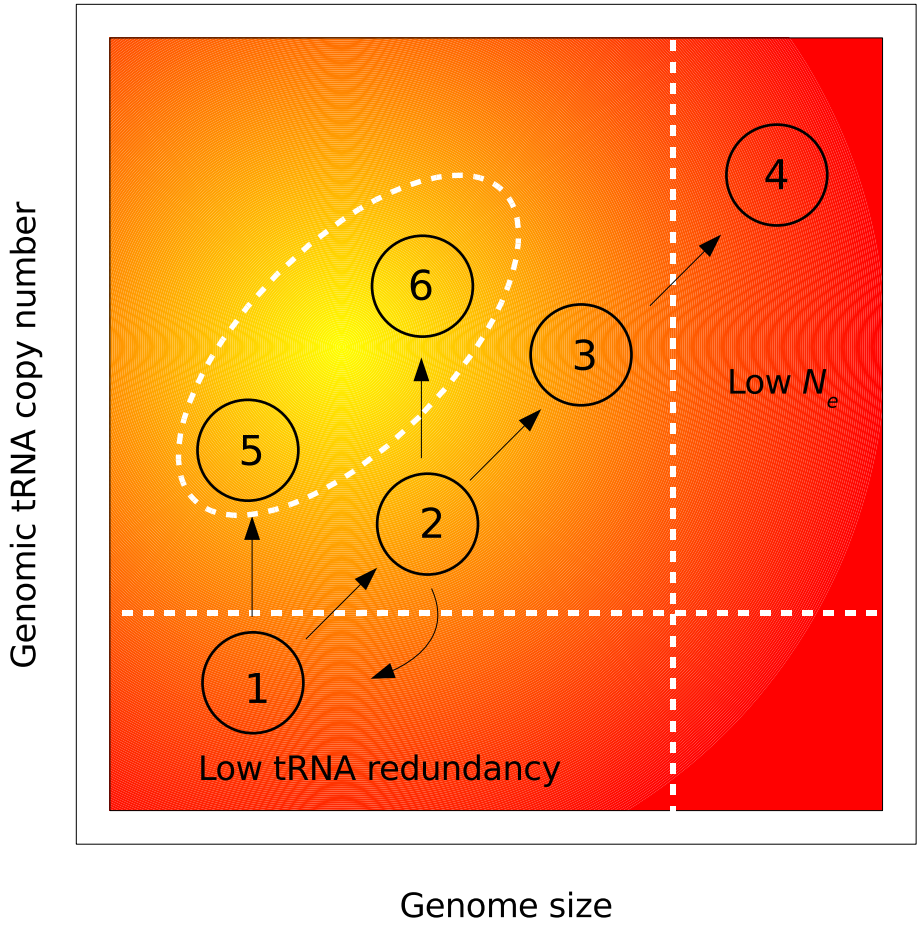

Figure 7.5: Hypothetical model of codon usage evolution.
A small genome sized ancestor (1), suffered a series of genome expansions (2-4). During this evolutionary process, the phylogeny would move into, and then out of a region where translational selection is operative. Selection for fast growth rate could lead to an increase in tRNA numbers without increase in genome size (5, 6), and the organisms would move into a selection hot spot (slanting ellipse). As genome size becomes larger, the concomitant reduction in population size causes a reduction in selected codon bias. The process could be reverted if selection for genome size or tRNA set reduction is present, such as in non-free living organisms such as *Plasmodium falciparum* or *Encephalitozoon cuniculi*.

7.3 Concluding remarks

It is now possible to trace the enigma of codon usage down to the evolution of tRNA genes and genome size and organisation. In organisms with small population sizes, the action of natural selection at the silent sites of codons is greatly impaired, these organisms have codon usage trends that are mainly determined by the particular mutational patterns of their genomes. In organisms with large population sizes, codon preferences are then determined by a balance between mutational bias and adaptation to the intracellular tRNA pool. However, the particular codons that are preferred in organisms with selected codon bias depends on the particular structure of their genomic tRNA pool. This structure, on the other hand, might have arisen as a result of partially stochastic processes. However a very interesting question arises here, certain bacteria with slow growth rates do not show evidence of strongly selected codon bias. Further studies are necessary in order to understand whether in these organisms population size is also reduced, or if simply the lifestyle is the main determinant of codon usage. Historically, there has been a segregation in the analysis of codon usage in Eukaryotes and Prokaryotes, and the objective of this work was partially to show that there are indeed conspicuous trends that explain the different roles of selection and mutational biases across all living organisms, and that both, Eukaryotes and Prokaryotes can (and should) be treated under the same framework despite the large differences in lifestyle and ecological roles of these organisms. I believe that it is under such a unified framework that the final answers to the riddles of codon usage will come to light.

Appendix

Derivation of the expected value of Nc

The following derivation assumes that the content of A equals that of T (or U in the case of RNA), and the content of G equals that of C. But the A + T content might be different from the G + C content. It is also assumed that the number of codons making up a protein is infinite. In the following discussion x is the silent GC content of a particular genome, and the system is in equilibrium.

All amino acids belonging to the two synonymous family (SF2) are always encoded by an A or T ending codon *and* by a C or G ending codon. Let c_{1j}be the codon ending in C or G, and c_{2j} the codon ending in A or T for the j-th amino acid in SF2. Then, the relative frequencies of c_{1j} and c_{2j} are simply x and $1-x$ respectively. Thus the expected homozygosity for the j-th amino acid is

$$\mathrm{E}(F_j) = x^2 + (1-x)^2,$$

and the expected Nc value for this amino acid is

$$\mathrm{E}(\mathrm{Nc}_j) = \frac{1}{x^2 + (1-x)^2}.$$

By definition, the contribution of the two synonymous family to the overall value of Nc is $\mathrm{C}_2 = 9/\bar{F}_2$ (see equation 3.3 on page 46). Because $\mathrm{E}(\mathrm{Nc}_i) = \mathrm{E}(\mathrm{Nc}_j)$ for all

i, j belonging to SF2, then $\mathrm{E}(\bar{F}_2) = x^2 + (1-x)^2$ and

$$\mathrm{E}(\mathrm{C}_2) = \frac{9}{x^2 + (1-x)^2}.$$

The only amino acid encoded by three codons is Isoleucine. These codons are ATT, ATC and ATA. The non normalised frequencies of these three codons are $(1-x)/2$, $x/2$ and $(1-x)/2$, since these are the frequencies of the respective ending nucleotides (T, C and A). We must normalise the frequencies so they add up to one. We divide them by their sum $(2-x)/2$ to get $(1-x)/(2-x)$, $x/(2-x)$, and $(1-x)/(2-x)$ respectively. Thus, the expected homozygosity is

$$\mathrm{E}(F_3) = 2\left(\frac{1-x}{2-x}\right)^2 + \left(\frac{x}{2-x}\right)^2,$$

and the expected Nc value for Isoleucine is

$$\mathrm{E}(\mathrm{Nc}_3) = \frac{(2-x)^2}{2(1-x)^2 + x^2}.$$

Because there is only one amino acid in SF3 then $\mathrm{E}(\mathrm{C}_3) = \mathrm{E}(\mathrm{Nc}_3)$

For amino acids belonging to the four synonymous family (SF4) the situation is very similar to the SF2 case. All codons encoding an amino acid belonging to SF4 are of the form XYT, XYC, XYA and XYG (*i.e.* the first two nucleotides are identical for the given amino acid), so the relative frequency of each codon is simply the frequency of the respective ending nucleotide. Thus, the expected homozygosity for amino acid j is

$$\mathrm{E}(F_j) = \frac{x^2 + (1-x)^2}{2},$$

and the expected Nc value for this amino acid is

$$\mathrm{E}(\mathrm{Nc}_j) = \frac{2}{x^2 + (1-x)^2}.$$

By definition, the contribution of the four synonymous family to the overall value of Nc is $C_4 = 5/\bar{F}_4$ (equation 3.3 on page 46). Because $E(Nc_i) = E(Nc_j)$ for all i, j belonging to SF4, then $E(\bar{F}_2) = [x^2 + (1-x)^2]/2$ and

$$E(C_4) = \frac{10}{x^2 + (1-x)^2}.$$

For amino acids belonging to the six synonymous family (SF6) the situation is more complex. There are three amino acids belonging to SF6, namely Serine, Leucine and Arginine. Serine presents the simplest case and we shall analyse it first. Leucine and Arginine behave similarly and will be treated together. Serine is coded for by the following six codons, shown below together with their normalised frequencies

$$\begin{array}{ll} \text{TCT} & (1-x)/3 \\ \text{TCC} & x/3 \\ \text{TCA} & (1-x)/3 \\ \text{TCG} & x/3 \\ \text{AGT} & (1-x)/3 \\ \text{AGC} & x/3. \end{array}$$

Thus the expected homozygosity for Serine codons is $E(F_{\text{Ser}}) = [(1-x)^2 + x^2]/3$. Below we show a similar table for Leucine with non normalised frequencies. Note that two of the codons start with T and four start with C, so the non normalised frequencies are the product of the frequencies of the first and third nucleotides.

$$\begin{array}{ll} \text{TTA} & (1-x)^2 \\ \text{TTG} & x(1-x) \\ \text{CTT} & x(1-x) \\ \text{CTC} & x^2 \\ \text{CTA} & x(1-x) \\ \text{CTG} & x^2. \end{array}$$

We know normalise to obtain

$$\begin{array}{ll} \text{TTA} & (1-x)^2/(1+x) \\ \text{TTG} & x(1-x)/(1+x) \\ \text{CTT} & x(1-x)/(1+x) \\ \text{CTC} & x^2/(1+x) \\ \text{CTA} & x(1-x)/(1+x) \\ \text{CTG} & x^2/(1+x), \end{array}$$

and the expected homozygosity for Leucine codons is $\mathrm{E}(F_{\mathrm{Leu}}) = 3\left(\frac{(x-x^2)^2}{(1+x)^2}\right) + 2\left(\frac{x^4}{(1+x)^2}\right) + \frac{(1-x)^4}{(1+x)^2}$. The case for Arginine is identical so $\mathrm{E}(F_{\mathrm{Arg}}) = \mathrm{E}(F_{\mathrm{Leu}})$. Then the expected average homozygosity for SF6 is $\bar{F}_6 = (\mathrm{E}(F_{\mathrm{Ser}}) + \mathrm{E}(F_{\mathrm{Leu}}) + \mathrm{E}(F_{\mathrm{Arg}}))/3$ and since by definition $\mathrm{C}_6 = 3/\bar{F}_6$ then

$$\mathrm{E}(\mathrm{C}_6) = \frac{27}{6\frac{3(x-x^2)^2+2x^2+(1-x)^4}{(1+x)^2} + x_g^2 + (1-x_g)^2}.$$

So for an arbitrary gene g, its expected Nc value $\mathrm{E}(\mathrm{Nc}_g)$ is simply the sum of all the expected contributions

$$\begin{aligned} \mathrm{E}(\mathrm{Nc}_g) &= \mathrm{E}(2+\mathrm{C}_2+\mathrm{C}_3+\mathrm{C}_4+\mathrm{C}_6) \\ &= 2+\mathrm{E}(\mathrm{C}_2)+\mathrm{E}(\mathrm{C}_3)+\mathrm{E}(\mathrm{C}_4)+\mathrm{E}(\mathrm{C}_6) \\ &= 2+\frac{19}{x_g^2+(1-x_g)^2}+\frac{(2-x_g)^2}{2(1-x_g)^2+x_g^2} \\ &\quad +\frac{27}{6\frac{3(x-x^2)^2+2x^2+(1-x)^4}{(1+x)^2}+x_g^2+(1-x_g)^2}. \end{aligned}$$

Bibliography

[1] Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, et al., 2000. The genome sequence of Drosophila melanogaster. Science 287:2185–95

[2] Akashi H, 1994. Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. Genetics 136:927–35

[3] Akashi H, 1995. Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in Drosophila DNA. Genetics 139:1067–76

[4] Akashi H, 1997. Codon bias evolution in Drosophila. Population genetics of mutation-selection drift. Gene 205:269–78

[5] Akashi H, 1999. Within- and between-species DNA sequence variation and the 'footprint' of natural selection. Gene 238:39–51

[6] Akashi H, 2001. Gene expression and molecular evolution. Curr Opin Genet Dev 11:660–6

[7] Akashi H, 2003. Translational selection and yeast proteome evolution. Genetics 164:1291–303

[8] Andersson SG, Kurland CG, 1990. Codon preferences in free-living microorganisms. Microbiol Rev 54:198–210

[9] Bachellier S, Gilson E, Hofnung M, Hill CW. Escherichia coli and Salmonella: cellular and molecular biology, chap. Repeated sequences. ASM Press, Washington, DC., pp. 2708–2720

[10] Bennetzen JL, Hall BD, 1982. Codon selection in yeast. J Biol Chem 257:3026–31

[11] 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. Proc Natl Acad Sci U S A 99:9697–702

[12] Bulmer M, 1987. Coevolution of codon usage and transfer RNA abundance. Nature 325:728–30

[13] Bulmer M, 1991. The selection-mutation-drift theory of synonymous codon usage. Genetics 129:897–907

[14] C elegans Sequencing Consortium, 1998. Genome sequence of the nematode C. elegans: a platform for investigating biology. Science 282:2012–8

[15] Carbone A, Zinovyev A, Kepes F, 2003. Codon adaptation index as a measure of dominating codon bias. Bioinformatics 19:2005–15

[16] Chamary JV, Hurst LD, 2004. Similar rates but different modes of sequence evolution in introns and at exonic silent sites in rodents: evidence for selectively driven codon usage. Mol Biol Evol 21:1014–23

[17] Chamary JV, Parmley JL, Hurst LD, 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat Rev Genet 7:98–108

[18] Charlesworth B, Sniegowski P, Stephan W, 1994. The evolutionary dynamics of repetitive DNA in eukaryotes. Nature 371:215–20

[19] Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P, 2006. Toward automatic reconstruction of a highly resolved tree of life. Science 311:1283–7

[20] Clarke B, 1970. Darwinian evolution of proteins. Science 168:1009–11

[21] Coghlan A, Wolfe KH, 2000. Relationship of codon bias to mRNA concentration and protein length in Saccharomyces cerevisiae. Yeast 16:1131–45

[22] Comeron J, Aguade M, 1998. An evaluation of measures of synonymous codon usage bias. J Mol Evol 47:268–74

[23] Comeron JM, 2004. Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence. Genetics 167:1293–304

[24] Crick FH, 1966. Codon–anticodon pairing: the wobble hypothesis. J Mol Biol 19:548–55

[25] Crow JF, Kimura M, 1970. An Introduction to Population Genetics Theory. Harper and Row

[26] Cunningham C, Omland K, Oakley T, 1998. Reconstructing ancestral character states: a critical reappraisal. Trends Ecol Evol 13:361–366

[27] Doolittle RF, Feng DF, Tsang S, Cho G, Little E, 1996. Determining divergence times of the major kingdoms of living organisms with a protein clock. Science 271:470–7

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

[29] dos Reis M, Wernisch L, 2009. Estimating translational selection in eukaryotic genomes. Mol Biol Evol 26:451–61

[30] 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 31:6976–85

[31] Drummond DA, Wilke CO, 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134:341–52

[32] Dunn KA, Bielawski JP, Yang Z, 2001. Substitution rates in Drosophila nuclear genes: implications for translational selection. Genetics 157:295–305

[33] 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 Genet 16:287–9

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

[35] Eddy SR, Durbin R, 1994. RNA sequence analysis using covariance models. Nucleic Acids Res 22:2079–88

[36] Edgar RC, 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113

[37] Edgar RC, 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–7

[38] Eyre-Walker A, Bulmer M, 1993. Reduced synonymous substitution rate at the start of enterobacterial genes. Nucleic Acids Res 21:4599–603

[39] Eyre-Walker A, Bulmer M, 1995. Synonymous substitution rates in enterobacteria. Genetics 140:1407–12

[40] Eyre-Walker A, Keightley PD, 2007. The distribution of fitness effects of new mutations. Nat Rev Genet 8:610–8

[41] Eyre-Walker AC, 1991. An analysis of codon usage in mammals: selection or mutation bias? J Mol Evol 33:442–9

[42] Felsenstein J, 1985. Phylogenies and the comparative method. Amer Nat 125:1–15

[43] Freeland SJ, Hurst LD, 1998. The genetic code is one in a million. J Mol Evol 47:238–48

[44] Fuglsang A, 2003. The effective number of codons for individual amino acids: some codons are more optimal than others. Gene 320:185–90

[45] Fuglsang A, 2004. The 'effective number of codons' revisited. Biochem Biophys Res Commun 317:957–64

[46] Fuglsang A, 2006. Estimating the "effective number of codons": the Wright way of determining codon homozygosity leads to superior estimates. Genetics 172:1301–7

[47] Garcia-Vallve S, Romeu A, Palau J, 2000. Horizontal gene transfer in bacterial and archaeal complete genomes. Genome Res 10:1719–25

[48] Goodenbour JM, Pan T, 2006. Diversity of tRNA genes in eukaryotes. Nucleic Acids Res 34:6137–46

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

[50] Grantham R, Gautier C, Gouy M, Mercier R, Pave A, 1980. Codon catalog usage and the genome hypothesis. Nucleic Acids Res 8:r49–r62

[51] Hale RS, Thompson G, 1998. Codon optimization of the gene encoding a domain from human type 1 neurofibromin protein results in a threefold improvement in expression level in Escherichia coli. Protein Expr Purif 12:185–8

[52] Harpending HC, Batzer MA, Gurven M, Jorde LB, Rogers AR, Sherry ST, 1998. Genetic traces of ancient demography. Proc Natl Acad Sci U S A 95:1961–7

[53] Hartl DL, Clark AG, 1997. Principles of population genetics. Sinauer Associates, Sunderland, Massachusetts, 3rd edn.

[54] Hartl DL, Moriyama EN, Sawyer SA, 1994. Selection intensity for codon bias. Genetics 138:227–34

[55] Hastie TJ, Tibshirani RJ, 1990. Generalized additive models. Chapman and Hall

[56] Hayashi T, Makino K, Ohnishi M, Kurokawa K, Ishii K, Yokoyama K, Han CG, Ohtsubo E, Nakayama K, Murata T, et al., 2001. Complete genome sequence of enterohemorrhagic Escherichia coli O157:H7 and genomic comparison with a laboratory strain K-12. DNA Res 8:11–22

[57] Holstege FC, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, Green MR, Golub TR, Lander ES, Young RA, 1998. Dissecting the regulatory circuitry of a eukaryotic genome. Cell 95:717–28

[58] Huelsenbeck JP, Bollback JP, 2001. Empirical and hierarchical Bayesian estimation of ancestral states. Syst Biol 50:351–66

[59] Huelsenbeck JP, Ronquist F, 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–5

[60] 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

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

[62] Ikemura T, 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 2:13–34

[63] Inokuchi H, Yamao F, 1995. tRNA: structure, biosynthesis and function., chap. Structure and expression of prokaryotic tRNA genes. ASM press, pp. 17–30

[64] Jansen R, Bussemaker HJ, Gerstein M, 2003. Revisiting the codon adaptation index from a whole-genome perspective: analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models. Nucleic Acids Res 31:2242–51

[65] Jukes TH, 2000. The neutral theory of molecular evolution. Genetics 154:956–8

[66] Jukes TH, Holmquist R, 1972. Evolution of transfer RNA molecules as a repetitive process. Biochem Biophys Res Commun 49:212–6

[67] Jukes TH, Osawa S, 1993. Evolutionary changes in the genetic code. Comp Biochem Physiol B 106:489–94

[68] Kanaya S, Yamada Y, Kinouchi M, Kudo Y, Ikemura T, 2001. Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with

translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. J Mol Evol 53:290–8

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

[70] Karlin S, Mrazek J, 1996. What drives codon choices in human genes? J Mol Biol 262:459–72

[71] Karlin S, Mrazek J, 2000. Predicted highly expressed genes of diverse prokaryotic genomes. J Bacteriol 182:5238–50

[72] Kimura M, 1962. On the probability of fixation of mutant genes in a population. Genetics 47:713–9

[73] Kimura M, 1968. Evolutionary rate at the molecular level. Nature 217:624–6

[74] Kimura M, 1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111–120

[75] Kimura M, 1981. Possibility of extensive neutral evolution under stabilizing selection with special reference to nonrandom usage of synonymous codons. Proc Natl Acad Sci U S A 78:5773–7

[76] Kimura M, 1983. The neutral theory of molecular evolution. Cambridge University Press

[77] Kimura M, Crow JF, 1964. The number of alleles that can be mantained in a finite population. Genetics 49:725

[78] King JL, Jukes TH, 1969. Non-Darwinian evolution. Science 164:788–98

[79] Koonin EV, 2005. Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet 39:309–38

[80] Koonin EV, Wolf YI, 2008. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucleic Acids Res 36:6688–719

[81] Kunst F, Ogasawara N, Moszer I, Albertini AM, Alloni G, Azevedo V, Bertero MG, Bessieres P, Bolotin A, Borchert S, et al., 1997. The complete genome sequence of the gram-positive bacterium Bacillus subtilis. Nature 390:249–56

[82] Kurland CG, 1993. Major codon preference: theme and variations. Biochem Soc Trans 21:841–6

[83] Lafay B, Atherton J, Sharp P, 2000. Absence of translationally selected synonymous codon usage bias in Helicobacter pylori. Microbiology 146:851–60

[84] Lafay B, Lloyd AT, McLean MJ, Devine KM, Sharp PM, Wolfe KH, 1999. Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases. Nucleic Acids Res 27:1642–9

[85] Lander E, Linton L, Birren B, Nusbaum C, Zody M, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al., 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921

[86] Lavner Y, Kotlar D, 2005. Codon bias as a factor in regulating expression via translation rate in the human genome. Gene 345:127–38

[87] Lawrence CB, McDonnell DP, Ramsey WJ, 1985. Analysis of repetitive sequence elements containing tRNA-like sequences. Nucleic Acids Res 13:4239–52

[88] Levy JP, Muldoon RR, Zolotukhin S, Link CJ Jr, 1996. Retroviral transfer

and expression of a humanized, red-shifted green fluorescent protein gene into human tumor cells. Nat Biotechnol 14:610–4

[89] Li W, 1987. Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. J Mol Evol 24:337–45

[90] Lowe TM, Eddy SR, 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25:955–64

[91] Lynch M, 2007. The origins of genome architecture. Sinauer Assoc.

[92] Lynch M, Conery J, 2003. The origins of genome complexity. Science 302:1401–4

[93] Lynn D, Singer G, Hickey D, 2002. Synonymous codon usage is subject to selection in thermophilic bacteria. Nucleic Acids Res 30:4272–7

[94] Man O, Pilpel Y, 2007. Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species. Nat Genet 39:415–21

[95] Man O, Sussman JL, Pilpel Y, 2007. Systems Biology and Regulatory Genomics, vol. 4023/2006 of *Lecture Notes in Computer Science*, chap. Examination of the tRNA Adaptation Index as a Predictor of Protein Expression Levels. Springer, pp. 107–118

[96] Marck C, Grosjean H, 2002. tRNomics: analysis of tRNA genes from 50 genomes of Eukarya, Archaea, and Bacteria reveals anticodon-sparing strategies and domain-specific features. RNA 8:1189–232

[97] Maside X, Lee A, Charlesworth B, 2004. Selection on codon usage in Drosophila americana. Curr Biol 14:150–4

[98] McClain WH, 1995. tRNA: structure, biosynthesis and function, chap. The tRNA identity problem: past, present and future. ASM Press, Washington, DC., pp. 335–347

[99] McInerney J, 1998. Replicational and transcriptional selection on codon usage in Borrelia burgdorferi. Proc Natl Acad Sci U S A 95:10698–703

[100] McVean GA, Charlesworth B, 1999. A population genetic model for the evolution of synonymous codon usage: patterns and predictions. Genet Res 74:145–48

[101] McVean GA, Charlesworth B, 2000. The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. Genetics 155:929–44

[102] Medigue C, Rouxel T, Vigier P, Henaut A, Danchin A, 1991. Evidence for horizontal gene transfer in Escherichia coli speciation. J Mol Biol 222:851–6

[103] Moszer I, Rocha E, Danchin A, 1999. Codon usage and lateral gene transfer in Bacillus subtilis. Curr Opin Microbiol 2:524–8

[104] Muto A, Andachi Y, Yuzawa H, Yamao F, Osawa S, 1990. The organization and evolution of transfer RNA genes in Mycoplasma capricolum. Nucleic Acids Res 18:5037–43

[105] Novembre JA, 2002. Accounting for background nucleotide composition when measuring codon usage bias. Mol Biol Evol 19:1390–4

[106] Ochman H, Lawrence J, 1996. Escherichia coli and Salmonella, Cellular and Molecular Biology, vol. II, chap. Phylogenetics and the amelioration of bacterial genomes. ASM Press, Washington D.C., pp. 2627–2637

[107] Ochman H, Wilson AC, 1987. Evolution in bacteria: evidence for a universal substitution rate in cellular genomes. J Mol Evol 26:74–86

[108] Orgel LE, Crick FH, 1980. Selfish DNA: the ultimate parasite. Nature 284:604–7

[109] Osawa S, Jukes TH, Watanabe K, Muto A, 1992. Recent evidence for evolution of the genetic code. Microbiol Rev 56:229–64

[110] Pagel M, 1999. Inferring the historical patterns of biological evolution. Nature 401:877–84

[111] Paradis E, Claude J, Strimmer K, 2004. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20:289–90

[112] Percudani R, Pavesi A, Ottonello S, 1997. Transfer RNA gene redundancy and translational selection in Saccharomyces cerevisiae. J Mol Biol 268:322–30

[113] Perriere G, Thioulouse J, 2002. Use and misuse of correspondence analysis in codon usage studies. Nucleic Acids Res 30:4548–55

[114] Petrov D, 2001. Evolution of genome size: new approaches to an old problem. Trends Genet 17:23–8

[115] Post LE, Nomura M, 1980. DNA sequences from the str operon of Escherichia coli. J Biol Chem 255:4660–6

[116] Post LE, Strycharz GD, Nomura M, Lewis H, Dennis PP, 1979. Nucleotide sequence of the ribosomal protein gene cluster adjacent to the gene for RNA polymerase subunit beta in Escherichia coli. Proc Natl Acad Sci U S A 76:1697–701

[117] Pupo GM, Lan R, Reeves PR, 2000. Multiple independent origins of Shigella clones of Escherichia coli and convergent evolution of many of their characteristics. Proc Natl Acad Sci U S A 97:10567–72

[118] Rasmussen CE, Williams CKI, 2006. Gaussian Processes for Machine Learning. The MIT Press

[119] Reid SD, Herbelin CJ, Bumbaugh AC, Selander RK, Whittam TS, 2000. Parallel evolution of virulence in pathogenic Escherichia coli. Nature 406:64–7

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

[121] Ronquist F, 2004. Bayesian inference of character evolution. Trends Ecol Evol 19:475–81

[122] Sawyer SA, Hartl DL, 1992. Population genetics of polymorphism and divergence. Genetics 132:1161–76

[123] Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J, Gibbs R, Hardison R, Miller W, 2000. PipMaker–a web server for aligning two genomic DNA sequences. Genome Res 10:577–86

[124] Sharp P, Bailes E, Grocock R, Peden J, Sockett R, 2005. Variation in the strength of selected codon usage bias among bacteria. Nucleic Acids Res 33:1141–53

[125] Sharp P, Stenico M, Peden J, Lloyd A, 1993. Codon usage: mutational bias, translational selection, or both? Biochem Soc Trans 21:835–41

[126] Sharp PM, 1991. Determinants of DNA sequence divergence between Escherichia coli and Salmonella typhimurium: codon usage, map position, and concerted evolution. J Mol Evol 33:23–33

[127] Sharp PM, Averof M, Lloyd AT, Matassi G, Peden JF, 1995. DNA sequence evolution: the sounds of silence. Philos Trans R Soc Lond B Biol Sci 349:241–7

[128] Sharp PM, Li WH, 1987. The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15:1281–95

[129] Sherry ST, Harpending HC, Batzer MA, Stoneking M, 1997. Alu evolution in human populations: using the coalescent to estimate effective population size. Genetics 147:1977–82

[130] Shields DC, Sharp PM, Higgins DG, Wright F, 1988. "Silent" sites in Drosophila genes are not neutral: evidence of selection among synonymous codons. Mol Biol Evol 5:704–16

[131] Stenico M, Lloyd AT, Sharp PM, 1994. Codon usage in Caenorhabditis elegans: delineation of translational selection and mutational biases. Nucleic Acids Res 22:2437–46

[132] Stoletzki N, Eyre-Walker A, 2006. Synonymous Codon Usage in Escherichia coli - Selection for Translational Accuracy. Mol Biol Evol

[133] Suyama M, Torrents D, Bork P, 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res 34:W609–12

[134] Tatusov RL, Koonin EV, Lipman DJ, 1997. A genomic perspective on protein families. Science 278:631–7

[135] Tautz D, 1999. Codon-preference riddles. Trends Genet 15:395

[136] Urrutia A, Hurst L, 2001. Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. Genetics 159:1191–9

[137] Urrutia AO, Hurst LD, 2003. The signature of selection mediated by expression on human genes. Genome Res 13:2260–4

[138] Venables WN, Ripley BD, 2002. Modern Applied Statistics with S. Springer

[139] Watanabe K, Osawa S, 1995. tRNA: structure, biosynthesis and function, chap. tRNA sequences and variation in the genetic code. ASM Press, pp. 225–250

[140] Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, et al., 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–62

[141] Watson JD, Crick FH, 1953. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 171:737–8

[142] Williams PD, Pollock DD, Blackburne BP, Goldstein RA, 2006. Assessing the accuracy of ancestral protein reconstruction methods. PLoS Comput Biol 2:e69

[143] Withers M, Wernisch L, Reis MD, 2006. Archaeology and evolution of transfer RNA genes in the Escherichia coli genome. RNA 12:933–42

[144] Wright F, 1990. The 'effective number of codons' used in a gene. Gene 87:23–9

[145] Xue H, Tong KL, Marck C, Grosjean H, Wong JT, 2003. Transfer RNA paralogs: evidence for genetic code-amino acid biosynthesis coevolution and an archaeal root of life. Gene 310:59–66

[146] Yang Z, 1994. Estimating the pattern of nucleotide substitution. J Mol Evol 39:105–11

[147] Yang Z, 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol 39:306–14

[148] Yang Z, 1996. Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol Evol 11:367–372

[149] Yang Z, 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555–6

[150] Yang Z, 2006. Computational Molecular Evolution. Oxford University Press, Oxford

[151] Yang Z, Kumar S, 1996. Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rates among sites. Mol Biol Evol 13:650–9

[152] Yang Z, Kumar S, Nei M, 1995. A new method of inference of ancestral nucleotide and amino acid sequences. Genetics 141:1641–50

[153] Yang Z, Nielsen R, 2008. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Mol Biol Evol 25:568–79

[154] Yang Z, Roberts D, 1995. On the use of nucleic acid sequences to infer early branchings in the tree of life. Mol Biol Evol 12:451–8

[155] Yokoyama S, Nishimura S, 1995. tRNA: structure, biosynthesis and function., chap. Modified nucleosides and codon recognition. ASM press, pp. 207–223