Evolution of Codon Usage in Bacteria

by

Ian Henry B.Sc.

School of Biology

Thesis submitted to the University of Nottingham for the Degree of Doctor of Philosophy, June 2007

Abstract

Initially, this thesis investigates patterns of intragenomic codon usage within the genome of the Delta Proteobacterium *Bdellovibrio bacteriovorus*. Correspondence analyses revealed the primary factor influencing codon usage within this genome to be related to translational selection. The relationship between the degree of codon usage adaptation (as given by the 'frequency of optimal codons' statistic) and putative gene expression level was used to look for genes with unusually high or low expression levels in *B. bacteriovorus*, in comparison to *Escherichia coli*, in order to gain further insight into the unusual lifestyle of this Delta Proteobacterium.

The scope was then broadened to explore intergenomic patterns of codon usage and initially extend a study measuring the strength of selected codon usage bias across bacterial genomes (Sharp *et al.* 2005). A dataset of 160 fully sequenced bacterial genomes was used and the strength of selected codon usage bias was seen to vary greatly between species. A correlation was observed between (log of) generation time and the strength of selected codon usage bias with fast growing bacteria showing a higher degree of selected codon usage bias than slow growing bacteria.

In bacterial species exhibiting significant levels of selected codon usage bias optimal codon choice was examined. It was observed that optimal codon choice is not always conserved across all bacterial genomes under selection but broad trends in optimal codon choice were seen to be associated with particular bacterial clades. In general, optimal codon choice was seen to be linked with differences in mutational biases among the clades, as seen by a correlation between optimal codon choice in particular clades and the G+C content of their genomes. Clades that were A+U rich (Firmicutes, Gamma Proteobacteria main clade) were seen to largely prefer codons of the form NNA/U whilst G+C rich clades (Alpha Proteobacteria, Actinobacteria and *Xanthomonas* species) showed preference for codons of the form NNG/C in their highly expressed genes.

Finally the relationship between optimal codon usage and tRNA abundances was explored. Changes in tRNA abundances were seen to coincide with switches in optimal codon usage. Therefore, switches in codon usage and tRNA abundance are thought to be influenced by changing mutational bias within the genome as reflected by the correlation between optimal codon choice, tRNA gene complements and genomic G+C content.

Acknowledgements

Firstly, I would like to thank my supervisor Paul for all his help and advice throughout my time here in Nottingham. I would also like to thank Eduardo Rocha for the bacterial generation time data used in this thesis. Liz Bailes and Louise Wain have also made coming to work enjoyable and have given me great support and kept me relatively sane over the past years and months.

The other PhD students in the department have made living and working in Nottingham a great time in my life with their coffee room and pub based shenanigans. I'd also like to thank friends outside the department including my housemates Jo and John as well as the members of Damage Litigation (a.k.a. Beeston Badgers or the Johnson Army) and SuperCleanUpTeam46 including Brian.

Primarily, though, I'd like to thank my parents for encouraging me throughout my whole education and putting it top in their list of priorities. They worked hard to allow me to achieve everything I have and I hope I've made their considerable efforts worthwhile. I'm sad that my dad was not able to see me complete my PhD due to his illness but before he died he made it clear that he knew I would achieve my PhD and that I should go on to achieve all that I wanted in life.

It is for these reasons that I'd like to dedicate this thesis to my dad, as without him I would never have come this far. He was a great man, always laughing and joking but with a caring serious side too. He'd do anything for me and when I was feeling down and ready to give up he'd always give me that energy to carry on. I will miss him greatly but know he'll be watching over me and I'm sure he'll be proud to see me complete my PhD.

This is for you dad, Edwin Henry, the late Bowling Champion of the World; with all my love.

Table of Contents

Table of Figures

Table of Tables

Chapter 1:
Introduction

1.1 Codon usage and the degenerate genetic code

The genetic code is degenerate as multiple codons can code for one amino acid. Such groups of codons coding for a single amino acid are known as synonymous codons. Some amino acids, such as serine, have as many as six synonymous codons whilst others are encoded by a single codon, which is the case for the amino acids methionine and tryptophan. In total 18 of the 20 amino acids can be encoded by more than one codon and most of this degeneracy is found at the third position in a codon. The groups of synonymous codons that encode for a particular amino acid are very well conserved over most species although a few small exceptions have been reported (Osawa *et al.*, 1992; Santos *et al.*, 2004).

Although one might expect synonymous codons to be used at approximately equal frequencies this is not the case in most bacterial genomes studied. Indeed, early work by Grantham and colleagues found that a certain consistency of codon choice is often found in genes of the same or similar genomes, with each genome having a 'system' for 'choosing' between codons (Grantham *et al.*, 1980a; Grantham *et al.*, 1980b); Grantham *et al*. termed this the 'genome hypothesis'. This observation was made using mRNAs from a variety of prokaryotic and eukaryotic species where it was seen that between mRNAs from different species there was degeneracy at the 3rd codon position (Grantham *et al.*, 1980a; Grantham *et al.*, 1980b). Correspondence analysis on the mRNA sequences showed a clustering of mRNAs with respect to genome type but no similar clustering was observed for the corresponding protein sequences encoded by each mRNA (Grantham *et al.*, 1980a); showing that genome specific codon usage was the main cause of variation in the mRNA.

1.2 Early work investigating patterns of codon usage in *E. coli*

Much of the early work on patterns of codon usage in bacteria was done in *E. coli*. Indeed the work done by Grantham and co-workers discussed above was done with *E. coli* as the main representative of bacteria, such was the availability of sequence data at the time.

Analysis of the ribosomal gene cluster adjacent to the RNA polymerase subunit β in *E. coli* found that codon usage in the ribosomal protein genes was highly non-random (Post *et al.*, 1979). In addition, it was noticed that codon preference was stronger in the ribosomal protein genes than in other genes. Further work, this time using DNA sequences for the *str* operon in *Escherichia coli* found that ribosomal protein genes and elongation factor genes in this operon had codon usage preferences corresponding to the most abundant isoaccepting tRNA species present in the cell (Post & Nomura, 1980). It was suggested that these non-random codon usage patterns that corresponded to tRNA abundances could be a result of a translational system adapted for translational efficiency, error minimization or both (Post & Nomura, 1980). Further work by Ikemura also found a relationship between patterns of codon usage and tRNA abundances in *E. coli* (Ikemura, 1981a; Ikemura, 1981b) and suggested that such codon usage patterns were a way of optimizing the process of translation. As a result codons selectively used in such a manner were defined as 'optimal' codons. Ikemura went on to state that codon choices in *E. coli* are constrained by tRNA availability and that this constraint is especially evident for highly expressed genes such as the ribosomal protein genes (Ikemura, 1981a; Ikemura, 1981b; Ikemura, 1985).

Correlations between gene expression levels and codon usage were also found by other researchers. Further work by Grantham on the 'genome hypothesis' incorporating new nucleic acid sequences made available subsequent to his 1980 papers (Grantham *et al.*, 1980a; Grantham *et al.*, 1980b) not only supported the initial 'genome hypothesis' but suggested

that codon choice was related to mRNA expressivity (Grantham *et al.*, 1981). The 29 bacterial genes (including 24 genes from *E.* coli) in the dataset were divided into 13 'highly expressed' genes and 16 'weakly expressed genes'. A correspondence analysis on these 29 mRNAs indicated strong modulation of coding strategy to messenger expression with 12 of the 13 highly expressed genes clustering distinctly from the weakly expressed genes (Grantham *et al.*, 1981).

Further analysis of 83 *E. coli* genes (Gouy & Gautier, 1982) took the observations regarding codon usage and tRNA abundances along with work looking at polypeptide elongation and tRNA cycling in *E. coli* (Gouy & Grantham, 1980) and concluded that highly expressed genes in *E. coli* use a subset of codons corresponding to the most abundant tRNA species so as to minimize the average number of tRNA discriminations per elongation cycle. This is based on the hypothesis that if tRNA species are present in high concentrations *in vivo* they are more likely to, by chance, interact with the ribosome at the A-site (Figure 1-1) (Gouy & Grantham, 1980). If an incorrect tRNA (one that does not match the codon to be translated) binds at the ribosomal A-site the aminoacyl-tRNA dissociates again from the ribosome, however when the specificity condition is fulfilled the elongation starts and transpeptidation and translocation occur. Gouy and Gautier used this to characterize a codon by the average number of codon-tRNA interactions at the A-site during one elongation cycle and stated that the concentration of the codon-cognate tRNA is equivalent to the probability of colliding with the A-site codon (Gouy & Gautier, 1982). The conclusion from this work was that genes expressed at high levels exhibit non-random codon usage in such a manner that codons corresponding to the most abundant tRNA species are used preferentially so as to minimize the average number of discriminations per elongation cycle. Thereby increasing the elongation rate and minimizing the chances of incorrect amino acid incorporation (Gouy & Gautier, 1982).

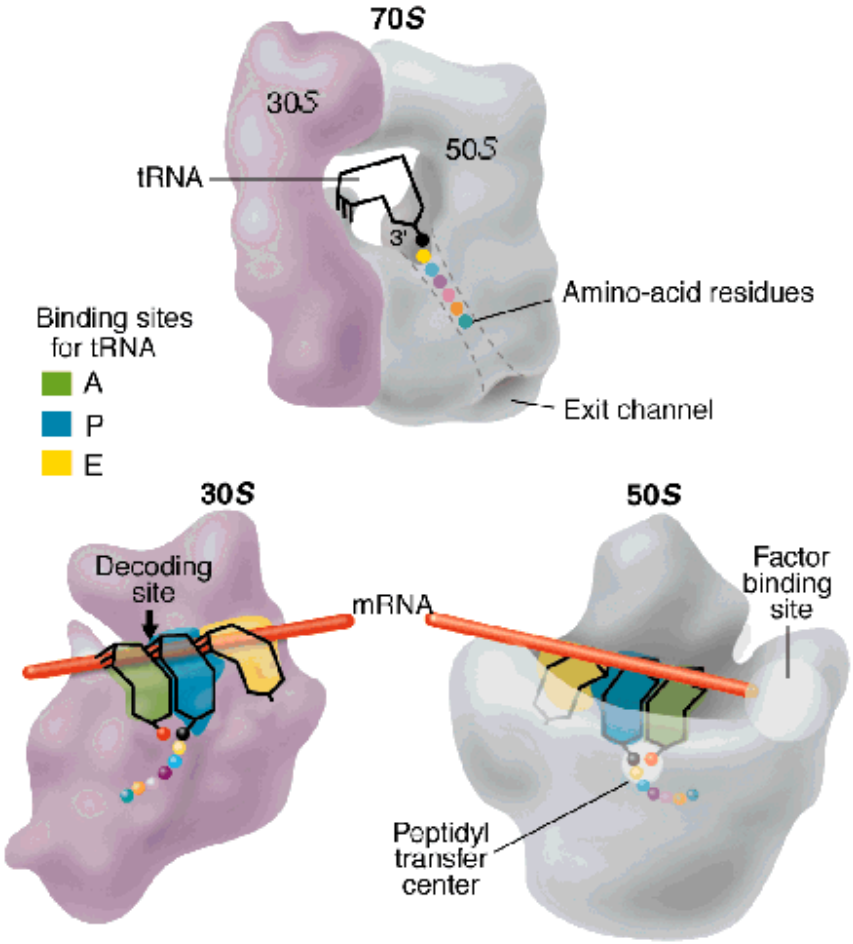

Figure 1-1 Structure of the bacterial 70S ribosome showing the ribosomal A, P and E tRNA binding sites
Aminoacyl-tRNA species enter the ribosome at the A-site and pair with the corresponding mRNA codon. Once this has occurred the ribosome moves one codon downstream and this shifts the tRNA, with its attached peptide, to the P site and opens the A site for the arrival of a new aminoacyl-tRNA. The E site holds a tRNA from which all peptide and amino acid has been removed (deacylated tRNA). This site is transiently occupied by the tRNA en route between leaving the P site and being released from the ribosome back into the cytosol. Taken from Liljas, 1999

Although this body of work began to make clear why codon usage in *E. coli* could be influenced by selection and highly co-adapted to tRNA abundances, it was less clear why certain synonymous codons were preferred over others. Work by Grosjean and Fiers, considering the nucleotide sequence of several highly expressed coding regions in bacteriophage MS2 as well as mRNAs from *E. coli*, suggested that the choice of optimal codon may be due to the requirement for an interaction of intermediate strength between codon and cognate tRNA at the A-site of the ribosome (Grosjean & Fiers, 1982). It was thought that overly strong or weak interactions were not favourable. Grantham, Gouy and Gautier agreed that codons with C and/or G nucleotides at the first and second positions should have a U at the third codon position due to the strong binding of C and G nucleotides and the weaker binding of A and U nucleotides (Grantham *et al.*, 1981). Similarly, they suggested that codons with A and/or U nucleotides at the first two codon positions should use C at the third position. Grantham noticed that pyrimidine ending codons did seem to be preferred in highly expressed genes in *E. coli* but noted that there was a contrast between amino acids encoded by two synonymous codons (duets), which mainly took the form NNC, and those with four synonymous codons (quartets) which mainly took the form NNU (Grantham *et al.*, 1981). He explained this by looking at the G+C composition of the first two codon positions in duets and quartets, noting that duets are largely A+U rich and so would be expect by the above theory to use mainly C at the third codon position. Similarly quartets are mainly G+C rich and so one would expect U to be used at the third codon position. In the case of duets with a choice of purine at the third position the pattern was seen to be less clear but the theory of maintaining a codon:anticodon interaction of intermediate strength was proposed for pyrimidine ending codons. However, such observations were made based on the limited sequence data available at the time and such predictions were a little ambitious, as discussed later in this thesis.

Codon usage for highly expressed genes, in *E. coli* at least, seemed to be related to optimizing translational accuracy and efficiency and was thought to be influenced by the abundance of cognate tRNA species. Although

highly expressed genes seemed to use codons corresponding to the most abundant cognate tRNAs it seemed the reverse was not true in that genes expressed at low levels did not intentionally use rare codons (Sharp & Li, 1986a). Instead, genes expressed at low levels in *E.coli* seemed to largely reflect dinucleotide frequencies in the genome as a whole (Sharp & Li, 1986a). This work also looked at genes known to be expressed at very high levels and those at moderately high levels and it was found that codon usage became increasingly more restricted the higher the gene expression level. It was therefore suggested that if the more highly expressed a gene was the more restricted its codon usage became then codon usage could be used as a way of predicting gene expression level. Sharp and Li developed a simple measure of synonymous codon usage bias which they termed the codon adaptation index (CAI) (Sharp & Li, 1987a). This index used a reference set of highly expressed genes to determine those codons that were used selectively in genes expressed at high levels. A score for a gene could then be calculated from the frequency of use of all codons in that gene. A similar scoring mechanism was also developed by Ikemura, which he termed the frequency of optimal codons (F_{OP}) (Ikemura, 1985).

1.3 Codon usage in bacterial genomes

Extensive study of codon usage in *E. coli* found that genes expressed at high levels had optimized codon usage patterns so that the codons used corresponded to the most abundant cognate tRNAs available. This observation led to the translational selection hypothesis under which selection could operate to optimize the translational machinery within the cell. This was done by having highly co-adapted optimal codons and cognate tRNA abundances to maximize the accuracy and efficiency of protein production within a cell and so confer a selective advantage to the bacterium. Translational selection was therefore implicated as a primary factor affecting codon usage in *E. coli*. Subsequent studies have found translational selection to influence codon usage in other bacterial genomes but two other major factors are also often seen to influence codon usage. These additional two factors are lateral gene transfer and strand bias.

1.3.1 Translational selection

The optimization of codon usage and tRNA abundances in species, such as *E. coli*, as part of translational selection was thought to be due to as many as three main factors; in order to maximize the speed of elongation, to maximize the accuracy of translation and to minimize the cost of proofreading.

Kurland suggested (reviewed in Kurland, 1991) that translational selection was brought about by the need to maximize translational elongation rate in genes expressed at high levels. He argued that such an increase in elongation rate would only be advantageous to genes expressed at high levels with mRNAs present in very high concentrations within the cell. He went on to state that in a cell processing a high number of different mRNA species, with most mRNAs showing unbiased codon usage, changing the rate at which one is translated by means of using optimal codons, would not increase the elongation rate. This is because, without such adapted codon usage patterns, after the ribosome completes the translation of one mRNA it is most likely to be sequestered by an entirely different mRNA species. In contrast, under optimal growth conditions where a greater proportion of metabolic activity is devoted to translation than under non-optimal conditions and a small number of different mRNA species are being processed, it is possible that the dominant group of mRNAs being translated on such occasions would be able to increase the speed of translation by using a very biased subset of codons. Under these circumstances ribosome number can become the limiting factor in translation, and because the availability of ribosomes to start new polypeptide chains is influenced by the speed at which they can complete the transit of mRNA (Andersson & Kurland, 1990), optimal use of ribosomes will maximize growth rate. Additionally, Kurland stated that, at maximal growth rates, optimization of translational efficiency is achieved when the concentration of tRNA species corresponding to the translated codons is maximized whilst at the same time the abundances of other tRNA species are minimized and so tRNA abundances should correspond to the restricted codon usage of the major proteins and vice-versa, as indeed seems to be the case in *E. coli* (Ikemura,

1981a; Ikemura, 1981b). Such a hypothesis as Kurland's would also predict that bacteria with a requirement for rapid growth are more likely to exhibit such patterns of codon usage within their genomes.

As well as the efficiency of translation it is also possible that accuracy of translation is a factor in translational selection. It was noticed that 'optimal' codons were mostly exactly complementary to the most abundant tRNA species in the cell (Gouy & Gautier, 1982; Ikemura, 1981a; Ikemura, 1981b). This ensures perfect Crick-Watson base pairing so as to improve translational accuracy. Such a reduction of translational misincorporation rates should confer a fitness advantage to the use of optimal codons. In addition if accuracy were important one would expect selection to minimize translational misincorporations, by the use of strongly restricted codon usage bias, at codons where the incorporation of an incorrect amino acid could result in the synthesis of a costly dysfunctional peptide. Patterns such as this have indeed been seen in *Drosophila melanogaster* (Akashi, 1994), where higher than usual optimal codon usage is evident in functionally important DNA-binding motifs as compared to other regions of the transcription factor genes, although this was not observed in *E. coli* (Hartl *et al.*, 1994). Recent work by Stoletzki and Eyre-Walker has shown, in contrast to the findings of Hartl, that optimal codons occur significantly more frequently at codons in which the amino acid is conserved than at non-conserved sites within the same gene (Stoletzki & Eyre-Walker, 2007). This discrepancy between the two studies is put down largely to the use of *Salmonella* as the comparison species by Hartl as compared to other *E. coli* strains (0157:H7 and CFT073) in the Eyre-Walker study (Stoletzki & Eyre-Walker, 2007). It is argued that between *E. coli* and *Salmonella* many amino acid substitutions are due to adaptive evolution and not random genetic drift as is more likely between *E. coli* strains. Similarly, if accuracy were important one may expect that among genes with similar expression levels selection to reduce translational misincorporations should be stronger in longer genes, because the cost of producing dysfunctional peptides should be proportional to their length (Eyre-Walker, 1996). Recent work by Eyre-Walker and colleagues confirmed this result in *E. coli* (Stoletzki & Eyre-Walker, 2007).

During the process of protein synthesis a ribosome must wait at a particular codon for the arrival of its complementary aminoacyl-tRNA. Ribosomes themselves are particularly costly to synthesize and so the time that they are not performing their function should be kept to a minimum (Akashi & Eyre-Walker, 1998). When the tRNA is bound to the mRNA in the ribosome GTP is hydrolysed which results in the release of elongation factor Tu. This kinetic proofreading step allows extra time to assess that the correct tRNA is bound, but this step is also quite costly. This is because if the wrong tRNA is found to be bound the tRNA has to be removed and recharged with EF-Tu and GTP once more. Even though proof reading mechanisms are present to ensure the correct translation of the mRNA, occasionally incorrect amino acids can be incorporated into the resultant polypeptide chain. Sometimes this will result in a functionless protein or a protein with reduced activity. Problems can also result from processivity errors while the ribosome is translocating, resulting in frameshift mutation or premature termination of the polypeptide chain. Once again these errors are most likely to produce functionless proteins and are likely to be detrimental.

Modification of tRNA species with regard to their structure or individual nucleosides may affect their affinity for particular codons. Nucleosides at positions 34 and 37 in the tRNA are often altered to modulate codon specificity (Santos *et al.*, 2004). This can be seen in *Mycoplasma* spp where a modified uracil at position 34 (Figure 1-2) prevents ambiguous decoding that before modification resulted from extreme wobble rules allowing the uracil to pair with A, C, G, or U at the third codon position (Yarian *et al.*, 2002). In addition tRNA modifications, particularly at position 37, have been implicated in reading frame maintenance by preventing slippage during translation (Agris, 2004). So it appears that modified tRNAs also contribute to translational accuracy and efficiency in addition to their influence over codon usage with a tRNA modification potentially altering codon preference.

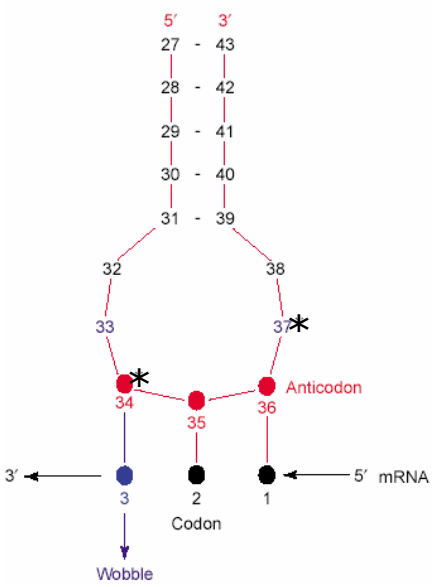

Figure 1-2 Cartoon showing tRNA anticodon structure.
Residues marked with a star are frequently modified nucleosides. Taken from Santos *et al*. 2004

1.3.2 Mutation-Selection-Drift

Studies of codon usage in genomes such as *E. coli* found that genes expressed at high levels had codon usage biased to the use of optimal codons maintained by selection to ensure translational efficiency. Factors affecting codon usage in weakly expressed genes were a matter of some debate.

At first, the theory that selection was acting on weakly expressed genes to regulate their expression by the use of minor codons was put forward (Grosjean & Fiers, 1982; Konigsberg & Godson, 1983). In contrast, others thought that codon usage patterns were brought about by a balance in a finite population between selection, which favours the incorporation of an optimal codon for each amino acid, and mutation along with drift, which allows non-optimal codons to persist within the population (Bulmer, 1991).

This theory of 'Selection-Mutation-Drift' (Bulmer, 1991) implies that selection is strongest in highly expressed genes so that these genes have strongly biased codon usage, but in weakly expressed genes there is a relaxation of selection and codon usage patterns in these genes are more susceptible to mutational pressures and genetic drift; as observed in *E. coli* (Sharp & Li, 1986a). Additionally, the synonymous substitution rate in genes expressed at low levels was found to be higher than in highly expressed genes as one would expect if selection pressure were relaxed (Sharp & Li, 1987b). It was also argued that a more efficient way to modulate gene expression would be to change the strength of the promoter or ribosome binding site (Sharp, 1986a; Bulmer, 1991).

Bulmer attempted to use the selection-mutation-drift hypothesis to create a population genetics based model. Selective forces likely to act on codon usage were evaluated and incorporated into this model to work out whether selection along with the effects of mutation rate and population size could produce the patterns of codon usage observed in bacterial genomes. Selective forces considered to be acting on codon usage by Bulmer were the

speed of translation, the accuracy of translation and the cost of proofreading; these factors were then incorporated into the model. The final model predicted the strength of selected codon usage bias in *E. coli* should be much higher than observed in reality. As Bulmer's model predicted selection should be so strong that no codons recognized by rare tRNAs should be observed in the genes sampled unless the effective population size were a factor of four less than the value used (10^5 not 10^9). Mathematically modeling the selective forces affecting codon usage patterns in a bacterial species was certainly an ambitious task and so it was perhaps unsurprising that such a model should not fit perfectly with reality.

Possible reasons why the model did not fit with observations were discussed by Bulmer (Bulmer, 1991). Firstly the problems involved an calculating the selection coefficient were numerous, with difficulties in calculating accurately the cost in terms of fitness of producing a dysfunctional protein, the possibility that the machinery involved in protein synthesis is regulated so as to buffer the impact of codon usage changes on fitness, as well as the relative impacts of selection for speed of translation, accuracy of translation, and the direct effect of accuracy of translation on errors in the protein product. Secondly, the effect of counterbalancing selection for the maintenance of mRNA secondary structure in opposition to the translational advantage of the most efficiently translated codon could reduce the strength of selection observed in reality. Thirdly, the model could be inadequate due to its failure to account for the genetic structure of clonal organisms, in particular the effect of selective sweeps through the population. An additional factor to consider was that of linkage. If there is linkage between sites, either because they are in the same gene or because the organism is clonal then selection is unable to act in a multiplicative way. For example, if two linked sites undergo a beneficial mutation in two separate bacteria then selection cannot possibly select for both mutations simultaneously, thereby reducing the effective population size (Li, 1987).

Despite the problems with the accuracy of the mathematical model produced the principles behind the model seem the most sensible.

Therefore it is believed that patterns of codon usage in bacterial genomes are due to the influence of selection and neutral mutation due to genetic drift. Patterns of codon usage therefore differ in individual genes as well as between bacterial genomes due to the differing strengths of selection and mutation.

1.3.3 Strand Bias

As well as the influence of translational selection, studies into the patterns of codon usage in bacterial genomes have often found strand bias to be a major factor influencing codon usage. However the extent to which this strand bias influences codon usage patterns does vary among bacterial genomes.

If no strand bias was present in a genome one would expect an equilibrium point where the base frequencies in each strand are always such that [A]=[T] and [G]=[C] regardless of the initial state of the DNA sequence or substitution patterns (Sueoka, 1995). If a significant deviation from the [A]=[T] and [G]=[C] condition is observed this is an indication that there is asymmetry in the substitution patterns of the two strands. Lobry's analyses of *Escherichia coli*, *Bacillus subtilis* and *Haemophilis influenzae* genomes showed that the substitution patterns of the two DNA strands in these genomes were asymmetric with the leading strand being more G+T rich than the lagging strand (Lobry, 1996). The unequal substitution patterns were seen to divide the bacterial chromosome into two segments with the deviation in substitution pattern switching at the origin and terminus of chromosome replication. Such a feature suggested that the asymmetry of substitution patterns may be linked to the replication and repair system of the organism. Indeed, asymmetry in the replication fork due to the anti-parallel nature of the strands and the enzymological asymmetry in the replication of the leading and lagging strand have been implicated as a possible cause of strand bias (Mrazek & Karlin, 1998).

Additionally a relative increase in bias in intergenic regions and at third codon positions showed that a relaxation of selective pressure increased the

bias suggesting some form of mutational bias (Lobry, 1996). Work by McLean and colleagues investigating 12 prokaryotic genomes, including 9 bacterial species, found evidence of a strong GC skew in eight of the nine bacteria, all of which switched at the probable origin and terminus of replication (McLean *et al.*, 1998). The pressure creating this GC skew appeared to be independent from the pressures determining genomic G+C content with genomes with low and high G+C contents showing evidence of strand bias.

The observation that, in most bacterial genomes, the majority of genes are found on the leading strand also led to the hypothesis that the asymmetry between the two DNA strands may be due to the effects of single strand deamination and transcription coupled repair (Francino & Ochman, 1997; Tillier & Collins, 2000; Francino & Ochman, 2001). Natural selection on codon usage does not appear to play a part in strand bias with the detection of substitutional bias in transcribed but untranslated regions as well as transcribed and translated regions (Francino & Ochman, 1997). The primary cause of strand bias is most probably down to replication biases but the effects of translation may, on occasion, add to this bias. Whatever the cause of strand bias, be it differences in strand replication and/or transcription, a neutral mutational process is thought to be responsible.

1.3.4 Lateral gene transfer

The effect of lateral gene transfer is another major factor that is seen to influence codon usage patterns in bacterial genomes. The significance of lateral gene transfer for bacterial evolution was first recognized in the 1950s when multidrug resistence patterns emerged on a worldwide scale. The ease with which certain bacteria were able to develop resistance to the same spectrum of antibiotics indicated that such traits were being transferred among taxa and not generated *de novo* by each lineage (Davies, 1996).

In order for lateral gene transfer to take place there must be a means by which a recipient cell can receive donor DNA. Once the donor DNA is in the

recipient cell the donor DNA must be incorporated into the recipient's genome or become associated with an autonomous replicating element. The incorporated genes must then be expressed suitably within the recipient cell. There are three main mechanism by which DNA can arrive in a recipient cell and integrate into the genome: transformation, transduction and conjugation (reviewed in Ochman *et al.*, 2000). Transformation is a process whereby naked DNA from the environment is taken up by a cell. Such a method of DNA transfer has the potential to transmit DNA between very distant organisms. Some species of bacteria may require specific recognition sequences in the DNA for effective uptake whilst others do not require such recognition sequences. The second method of DNA transfer is the process of transduction. This method requires the use of a bacteriophage to introduce DNA into a new bacterium. The bacteriophage must replicate within an organism and package either random DNA fragments, in the case of generalized transduction, or the DNA adjacent to the phage attachment site, in the case of specialized transduction. The quantity of DNA that can be transferred by a phage depends on the size of the phage capsid but is of the order of hundreds of kilobases. Phage transduction between organisms is limited to those with receptors recognized by the bacteriophage on their surface. However, the process can be more efficient as phage proteins can promote integration of DNA into the recipient chromosome and protect transferred DNA from degradation by the host. The third method by which lateral gene transfer can take place is that of conjugation. Conjugation typically involves the transfer of DNA by either a self-transmissible or mobilizable plasmid. Conjugation can also occur through the transfer of chromosomal sequences by plasmids that can integrate into a host chromosome. A further method of transfer by conjugation involves conjugative transposons, which encode proteins required for their excision from the donor, formation of a conjugative bridge and transposition into the recipient strain. Once the DNA has entered a host cell it can adopt several methods in order for it to be stably maintained with its new host. Persistence as an episome, homologous recombination, integration into the host chromosome mediated by bacteriophage integrases or mobile element transposases and illegitimate incorporation by chance through double strand break repair are

all methods that can enable a DNA sequence to be maintained in a new host.

Via the mechanisms outlined above almost any sequence can be transferred between organisms. However, it is suggested that prokaryotic genes can be divided into two functionally distinct gene classes, 'informational genes' and 'operational genes' (Rivera *et al.*, 1998). It is thought that informational genes, such as those involved in translation, transcription and replication are less easy to transfer than operational genes such as those involved in metabolism. A study by Nakamura and coworkers used a dataset of 116 prokaryotic genomes to look for evidence of lateral gene transfer (Nakamura *et al.*, 2004). They found that 14% of open reading frames in these 116 prokaryotes were the subject of recent horizontal transfer. Genes found to be horizontally transferred were found to be largely from three main gene categories: cell surface, DNA binding and pathogenicity. Pathogenicity-related genes were largely involved with toxin production or resistance. Genes involved with the cell surface were mostly involved with surface structure (e.g. fimbrial or pilus protein genes) or biosynthesis and degradation of surface polysaccharides and lipolysaccharides. These surface structure genes may also play a role in pathogenesis as they enable microbes to attach to host cells. DNA binding proteins can promote or inhibit transcription regulation and their role may be to alter gene expression patterns in a recipient organism. The next highest proportion of horizontally transferred genes was in the 'DNA metabolism' category and was mainly due to restriction/modification system (RMS) genes. These RMS genes are believed to be frequently involved in horizontal transfer events between organisms (Kobayashi, 2001).

Detecting such horizontal transfer events can be carried out by a variety of subtly different methods but all tend to rely on detecting unusual changes G+C content or patterns of codon usage within regions of a host genome. This is because at the time of introduction a recently transferred gene will have the codon usage pattern and base composition of the genome it was

transferred from. However, after transfer the gene becomes subject to the mutational processes affecting the recipient genome and so the sequence will incur substitutions until eventually it reflects the DNA composition of the new genome. Such a process of 'amelioration' is a function of the relative rate of G/C to A/T mutations (Lawrence & Ochman, 1997). This process means that the more recently a gene has been transferred the easier it is to find as its base composition will be more likely to be different from its new host.

Evidence of lateral gene transfer in *E. coli* was observed by Medigue and co-workers in 1991 in a dataset of 780 *E. coli* genes using correspondence analysis (Medigue *et al.*, 1991). The results of this analysis identified genes corresponding to surface elements of the cell, genes coming from mobile elements as well as gene resulting in a high fidelity of DNA replication which where all implicated in acquisition via horizontal transfer (Medigue *et al.*, 1991). Additional research by Lawrence and Ochman further investigated lateral gene transfer in the *E. coli* genome and found 755 of the 4288 open reading frames (ORFs) had been introduced into *E. coli* by at least 234 lateral transfer events since *E. coli*'s divergence from *Salmonella* 100 million years ago (Lawrence & Ochman, 1998), thus indicating the frequency of horizontal transfer between bacterial genomes.

1.4 Codon usage variation between bacterial genomes

Patterns of codon usage can vary widely between bacterial species. All of the factors discussed in the previous section have differing degrees of influence on the overall codon usage patterns depending on the bacterial species in question.

The strength of selected codon usage bias can vary dramatically between species with species such as *E. coli* showing strong evidence of translational selection, as discussed previously in this chapter, whilst other bacteria such

as *Helicobacter pylori* show little evidence for the influence of selection on their codon usage patterns (Lafay *et al.*, 2000). Even in bacteria under the influence of selection the genes most influenced by selection may differ. Whilst genes such as the ribosomal protein genes would be expected to show strong selected codon usage bias in all genomes where selection is a key factor other genes targeted by selection may vary according to the pressures on the bacterium in question. Indeed, dependent on niche and lifestyle, different genes may be expressed at high levels in different bacterial species. In addition to this the synonymous codons that are optimal may be not be the same in different bacterial species. For example, the genomes of *E. coli* and *Bacillus subtilis* are both influenced by selected codon usage bias but whereas *E. coli* prefers the CCG codon for the amino acid proline, *B. subtilis* prefers the alternative CCA codon.

The strength of strand bias can also be seen to vary between species. Bacteria such as *Borrelia burgdorferi* and *Treponema pallidum* have patterns of codon usage strongly influenced by strand bias whilst bacteria such as the S*ynechocystis* species exhibit little effects of strand bias (McLean *et al.*, 1998). Similarly the relevance of lateral gene transfer varies hugely between species with bacteria such as *Bradyrhizobium japonicum* and *Neisseria meningitidis* MC58 having an estimated 23.2% and 21.9% of their genomes due to horizontally transferred genes whilst genomes such as the *Rickettsia* and *Buchnera* species having less than 5% and as little as 0.5% of their genes due to horizontal transfer (Nakamura *et al.*, 2004).

All of these factors contribute, in differing degrees depending on the genome in question, to variation from the genome 'default' codon usage pattern. This default may also vary between species and is thought to be largely due to differing mutational biases between bacterial genomes

1.4.1 Variation in G+C content

Genomic G+C content values can range dramatically between bacterial species from extremely G+C rich genomes such as *Micrococcus luteus* (G+C

content: 72%) (Ohama *et al.*, 1990) to genomes with a low G+C content such as *Mycoplasma capricolum* (G+C content: 25%) (Ohkubo *et al.*, 1987).

Sueoka suggested that this variation in G+C content could be explained by 'directional mutational pressure' (Sueoka, 1962; Sueoka, 1988). This theory stated that the major cause for a change in DNA G+C content of an organism was the rate mutation between an α-pair (A-T or T-A) and a γ-pair (G-C or C-G). The wide ranging G+C contents among species were then explained by the differences in mutation rates from α to γ pairs and γ to α pairs. These relative mutation rates were thought to vary among bacterial species, leading to differing equilibrium positions and hence differing G+C contents. Support for this model came from the discovery that a mutator gene (*mutT*) in *E. coli* caused transversions from AT to CG pairs. Such a transversion event occurring with unidirectional preference was shown, over 1200-1600 generations, to be able to change G+C content by 0.2-0.5% (Cox & Yanofsky, 1967). It, therefore, seemed that directional mutational pressure could change the G+C content of a genome. Sueoka also argued that the effects of directional mutation pressure can be constrained by selective forces acting upon the genome. Thus, 3rd codon positions, where silent sites allow the effects of selection to be reduced, vary much more in G+C content than 1st and 2nd codon positions (Muto & Osawa, 1987; Sueoka, 1988). More recently, work by Chen and co-workers investigating patterns of codon usage in 100 bacterial and archaeal species found two parameters could differentiate the genome-wide codon bias of all these species (Chen *et al.*, 2004). The first correlated strongly with genomic G+C content and was deemed due to directional mutational pressure, whilst the second factor correlated with context-dependent nucleotide bias. The primary feature influencing G+C content and codon usage seemed to be made up of these two factors, with selective forces acting on translated sequences being only a secondary effect. Chen *et al*. showed that overall codon usage in prokaryotic genomes can be estimated by analysing intergenic sequences alone, thereby ignoring any selective forces acting on translated sequences within the genome.

In contrast to the theory of directional mutational pressure which explained compositional differences in genomes by neutral processes there were others who believed such variation in G+C content could be explained by the action of selection. This argument stated that environmental pressures can constrain genomic composition and affect both coding and non-coding sequences alike, such that all sequences in the genome together comprise a 'genome phenotype' influenced by the effects of natural selection (Bernardi & Bernardi, 1986). It was, therefore, suggested that bacteria living in hot conditions should have high genomic G+C contents, as favoured by selection.

Work by Galtier and Lobry found no such correlation between G+C content and optimal growth temperature although sequences with known secondary structure, such as tRNA and rRNA sequences did show a correlation between G+C content and optimal growth temperature (Galtier & Lobry, 1997). It was so concluded by Galtier and Lobry that it seemed unlikely that G+C content on a genome wide scale should be influenced by environmental pressures such as temperature. However, a series of papers by Musto in conjunction with Bernardi and others showed that correlations between genomic G+C content and optimal growth temperatures exist in prokaryotes (Musto *et al.*, 2004; Musto *et al.*, 2005; Musto *et al.*, 2006) at the close family level at least. This was interpreted as evidence based around the observation that G:C pairs are more stable than A:T pairs due to the extra hydrogen bond in G:C pairs, thus it would be advantageous for organisms living at high temperature to have a higher G+C content. Musto argued that a wide range of prokaryotes included in the dataset could introduce 'often contrasting inputs on genome composition', thus explaining why correlations were only observed at the family level. These results were, in turn, disputed by other researchers (Marashi & Ghalanbor, 2004; Wang *et al.*, 2006). Most researchers now seem to agree that genomic G+C content seems to be largely determined by neutral mutational processes.

1.4.2 Variation in the strength of selected codon usage bias

Individual studies concerning a wide variety of bacterial genomes have shown that the strength of selection present within a genome can vary widely. The classical example of a genome exhibiting selected codon usage bias is that of *E. coli* where it can be seen that highly expressed genes, such as the ribosomal protein genes, use a subset of optimal codons that correspond to the most abundant tRNAs in the cell (Ikemura, 1985). In contrast other genomes such as *Helicobacter pylori* have been shown to exhibit little selection if any (Lafay *et al.*, 2000).

It was hypothesized that the resulting codon usage in a genome was due to the effects of both selection and the combination of mutation and drift (Bulmer, 1991; Sharp & Li, 1986b). Genomes such as *E. coli* with moderate G+C content and strong selection show predominantly the effects of selected codon usage bias whilst genomes with weak selection have codon usage that mainly reflects the underlying mutational biases influencing the genome. There is also some evidence in genomes, such as *Streptomyces coelicolor*, of strong mutational biases driving codon usage to extremely high G+C content, thus masking the effect of selected codon usage bias (Wright & Bibb, 1992).

Why should some genomes exhibit strong selection and others relatively little selection? Recent studies have focused on trying to quantify the amount of selection present in many bacterial genomes in order to learn more about the factors influencing codon usage. Three studies have attempted to quantify the strength of selection in a wide variety of organisms using different approaches (dos Reis *et al.*, 2004; Rocha, 2004; Sharp *et al.*, 2005).

Dos Reis *et al.* looked at 126 fully sequenced genomes from archaea to eukaryotes, including 101 bacterial genomes. The technique used to test for translational selection employed two statistics. The first was based on

the effective number of codons in a gene (Wright, 1990) and was an attempt to look for restricted codon usage bias, i.e. whether a gene used synonymous codons randomly or deviated from totally random usage. The second statistic was termed the tRNA adaptation index, a modification of the codon adaptation index (CAI) of Sharp & Li (1987a). This statistic looked at tRNA gene copy number in a species (using this as a surrogate for tRNA levels in the cell) and combined this with the strength of codon-anticodon interaction to assign fitness values to codons. The tRNA adaptation index value for a gene was an attempt to estimate how adapted a particular gene's codon usage was to the tRNA pool available and was an average of the fitness values assigned to each codon present in the gene. Dos Reis *et al*. looked at the correlation of the two statistics, arguing that this was a measure of the strength of selected codon usage bias. The presence of restricted codon usage (statistic one) where codons used correspond to the most abundant tRNAs in the cell (statistic two) was interpreted as a sign of selected codon usage bias. This method found evidence of selection in 26% of the 101 bacterial genomes analyzed. The conclusions of dos Reis and co-workers were that translation selection is strongly influenced by the co-evolution of genome size and tRNA redundancy (dos Reis *et al.*, 2004).

Rocha also considered variations in codon usage bias from using tRNA abundance data (Rocha, 2004). This work involved the analysis of the tRNA gene pool of 102 bacterial species. The amount of general codon bias in each genome was estimated by comparing the effective number of codons (ENC') (Wright, 1990; Novembre, 2002) in ribosomal protein genes with the effective number of codons in the genome as a whole. Genomes with more restricted codon usage bias in their highly expressed ribosomal genes as compared to the genome as a whole were found to have more tRNA gene copies. Minimal generation times for these 102 species were also examined and it appeared that organisms with rapid generation time had genomes with more tRNA genes but less tRNA gene redundancy (i.e. fewer anticodon species) as well as more restricted codon usage bias. It was argued that this over-representation of some anticodon species

suggested an optimization of the translational machinery to use a subset of optimal codons corresponding to these overrepresented tRNA species.

Finally, work by Sharp and colleagues attempted to compare the strength of selection among 80 bacterial species (Sharp *et al.*, 2005). The method used here was based on the population genetics model of selection-mutation-drift as devised by Bulmer (Bulmer, 1991). This method looked at just 4 amino acids (Phe, Tyr, Ile, Asn) where optimal codon choice was conserved across all bacterial genomes examined. The strength of selected codon usage bias was based on how often the optimal synonymous codons for these four amino acids were used in a subset of 40 highly expressed genes as compared to the genome as a whole. This method was designed to take into account genomic G+C content and used a specific selected bias measure instead of the general bias measure (based on the effective number of codons) used by Rocha. Results showed that the strength of selected codon usage bias was strongly correlated to rRNA operon number (used as a surrogate for generation time) and also tRNA abundance.

All three methods found variation in the strength of selection across a wide variety of species. Sharp and Rocha concluded that the variation in selection was most probably due to difference in lifestyle and especially generation time of the different species. On the other hand, dos Reis put the difference in strength of selection down to genome size. Sharp argued that this conclusion was unjustified with regard to bacterial genomes and that the results of dos Reis *et al.* were heavily influenced by the inclusion of eukaryote species. Indeed Sharp found that 10 of the 11 species, in his 80 genome dataset, with >5000 genes had <75 tRNA genes, while 10 of the 11 species with >75 tRNA genes had <500 protein-coding genes (Sharp *et al.*, 2005). Therefore, the overall conclusions of these studies have implicated the rate at which a species can reproduce as an important factor, with organisms with a rapid generation time requiring faster and more efficient translational machinery.

It also seemed that the balance between mutation and selection could be swayed by effective population size. A small effective population size may enhance random genetic drift, and so inhibit codon selection. Indeed many parasitic species, such as the intracellular parasites of genera *Buchnera*, *Wigglesworthia* and *Rickettsia* where effective population size is thought to be small, showed low amounts of selection so that the major influences on codon usage appeared to be neutral features such as mutational bias (Andersson & Andersson, 1999; Mira & Moran, 2002; Moran, 1996; Sharp 2005). These genomes are expected to be relatively clonal with little recombination (low recombination has the same effect as low population size due to linkage of sites) but other species such as *E. coli* and *H. influenzae* exhibit high selected codon usage bias even with relatively low recombination rates, so clearly there are many factors influencing the strength of selection.

1.4.3 Variation in optimal codon choice

It has long been known that different bacterial species can have different optimal codons from each other. For example codon choice in *E. coli* and *B. subtilis* can be seen to be different with *E.coli* preferring the use of the CAG codon for the amino acid glutamine and CCG for the amino acid proline whilst *B. subtilis* shows preference for the CAA codon and CCA codon respectively. If the common ancestor of two species with adapted codon usage today, but showing different optimal codons, already had adapted codon usage, how did divergence in the identity of optimal codons occur?

It is possible that a relaxation in selective pressure, such as that caused by a population bottleneck could result in the loss of selected codon usage bias within a genome. Codon usage patterns would then be influenced by mutational biases within the genome without selective pressure ensuring the maintenance of specific restricted codon usage patterns. A change in directional mutational pressure (Sueoka, 1988) would cause codon usage patterns in the genome to slowly change to reflect the new mutational bias until an equilibrium position was reached. A re-establishment of selection at this point would result in restricted codon usage patterns once again, but

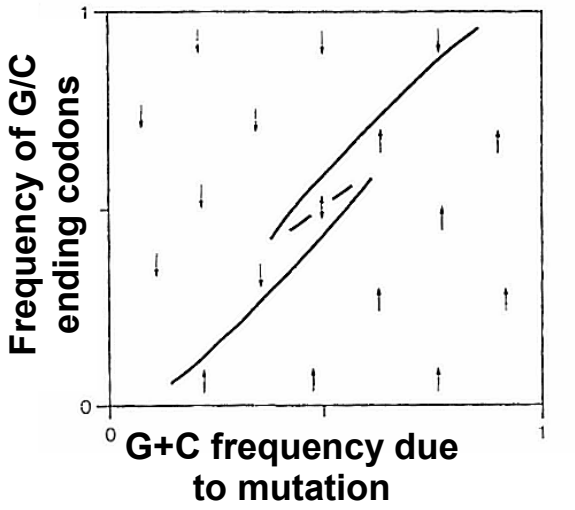

Figure 1-3 Plot Shields' of Shields' curve for genes under low selection

Shields predicted that in genes under low selection the frequency of G/C ending codons will change smoothly under the influence of a changing mutational bias. Continuous lines represtent equilibrium codon frequencies whilst broken lines represent unstable equilibrium frequencies. Arrows show the direction of movement of codon frequency at a given G+C frequency due to muatation. Figure modified from Shields, 1990.

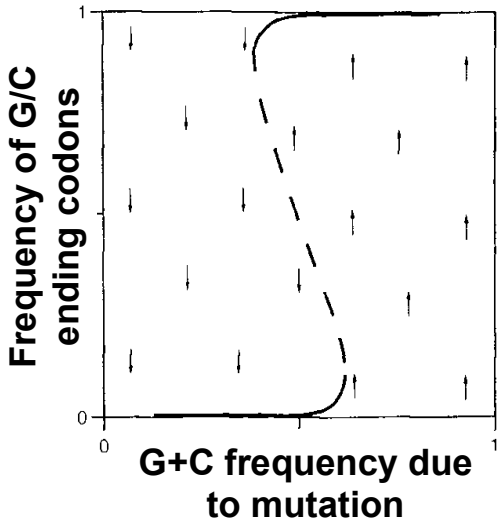

Figure 1-4 Plot Shields' of Shields' curve for genes under high selection

Shields predicted that in genes under high selection the frequency of G/C ending codons will switch sharply under the influence of a changing mutational bias. Continuous lines represtent equilibrium codon frequencies whilst broken lines represent unstable equilibrium frequencies. Arrows show the direction of movement of codon frequency at a given G+C frequency due to muatation. Figure modified from Shields, 1990.

this time derived from the new genomic G+C content with a new set of optimal codons and co-adapted tRNA abundances.

An alternative hypothesis was proposed by Shields (1990) whereby optimal codon usage could change without a relaxation in selection. Firstly he considered the case of genes under low selection. In lowly expressed genes the selective constraints on codon usage are low and so patterns of codon usage in such genes are largely determined by mutational biases (Figure 1-3). Under these circumstances if the mutation bias changes the equilibrium codon frequency position would move slowly to reflect this.

In contrast, highly expressed genes under strong selective constraints should resist the changing mutational bias and maintain their restricted optimal codon usage patterns. However, a shift in mutational bias over a critical range would result in a complete switch in preference from one codon to another. This is because the organism must translate the weakly expressed genes, which have taken on new codon usage patterns reflecting the change in mutational bias, as well as the highly expressed genes, and so eventually the existing optimal codon preferences can no longer be maintained against this strong change in mutational bias. This results in a sudden switch in preference. When the selection pressure is stronger, then a stronger mutational bias against a codon is required to switch its selective advantage. No stable equilibrium position in the middle is present where intermediate codon usage patterns are observed as this would not be advantageous to the organism (Figure 1-4).

Not only did Shields outline this model, he looked for evidence of such codon usage patterns in a variety of organisms. To see if the actual patterns of codon usage in highly expressed genes fit in with the pattern predicted, Shields plotted codon frequencies in highly expressed genes against mutational bias (estimated from codon frequencies in lowly expressed genes). One may expect that actual codon usages are at stable equilibria and so the theory would predict that the data points should lie on

the upper and lower arms of the S-shaped curve. Evidence of such patterns of codon usage did exist and were especially convincing for the amino acid lysine although the lack of sequence data meant that only seven genomes were analyzed (Shields, 1990). Shields additionally looked at codon preferences among the Enterobacteria for evidence of changes in codon preference. This study looked at just the three species *E. coli*, *Serratia marcescens* and *Proteus vulgaris*. The genomic G+C contents of these genomes were taken to be 51%, 59% and 39% respectively. Codon usage preference between *E. coli* and *S. marcescens* in highly expressed genes were remarkably similar whereas lowly expressed genes exhibited stronger G+C ending codon usage patterns reflecting the stronger G+C mutational bias present in the *S. marcescens* genome (Shields, 1990; Sharp, 1990). It therefore seemed that a stronger mutational bias, whilst not strong enough to alter codon usage in highly expressed genes, had influenced codon usage patterns in genes under low selection. In contrast the A+T mutational bias in the *P. vulgaris* genome appeared to have been able to cause changes to codon preference in highly expressed genes. The reasons why a switch should occur for *P. vulgaris* and not *S. marcescens* were unclear to Shields, however he suggested that larger effective population sizes in the *S. marcescens* lineage may have been enough to prevent mutational bias from causing a switch in optimal codon preference. Shields also pointed out that *P. vulgaris* diverged from *E. coli* more than twice as long ago as *S. marcescens* which may be significant as codon frequencies and the translational machinery may take a long time to gradually co-evolve.

1.5 Aims of this thesis

The first aim of this thesis was to investigate the nature of codon usage variation within the genome of a bacterial species that had not previously been examined. The species chosen was *Bdellovibrio bacteriovorus* (Rendulic *et al.*, 2004). This species belongs to the Delta Proteobacteria, and thus represented a bacterial phylum which had not previously been analyzed in this way. Also, it is a species with a very unusual lifestyle, since it preys on other Gram-negative bacteria such as *E. coli* and is able to enter the periplasm of such bacteria whereupon it degrades the contents of

the host cell and extracts the degraded cell material for use in growth and reproduction. Interest in *Bdellovibrio* has been great due to its ability to kill other pathogenic bacteria whilst being unable to infect mammalian cells (Lenz & Hespell, 1978); this has led to it being dubbed a 'living antibiotic'. Furthermore a colleague in this department, Professor Liz Sockett, was involved in the sequencing of the genome as well as being greatly experienced in the study of this bacterium. Therefore, it made sense to collaborate with her to investigate whether codon usage analysis could provide insights into the biology of this species. Additionally, in the process of this analysis I could become familiar with key techniques used in the analysis of codon usage bias within bacterial genomes, as well as addressing some issues that had been raised with regard to correspondence analysis methods (Perrière & Thioulouse, 2002).

The next aim of the thesis was to look at patterns of codon usage bias between bacterial genomes. Previous studies of individual bacterial genomes had shown that the strength of codon usage bias varies considerably between species. The accumulation of vast quantities of genomic sequence data allowed, for the first time, large scale studies of factors affecting codon usage bias across bacterial genomes. Three groups had attempted to look at such variation and quantify the strength of selection (dos Reis *et al.*, 2004; Rocha, 2004; Sharp *et al.*, 2005). The work in this thesis most closely follows on from the Sharp *et al.* (2005) work and its attempt to quantify the strength of selected codon usage bias in 80 bacterial genomes. The rapid rate at which genomic sequence data for bacterial species is accumulating allowed a similar study to be repeated at a greater resolution with a much larger dataset. The aim of this study was not only to look at the variation in selected codon usage bias but to try and understand the factors influencing the degree to which selection influences codon usage bias.

Thirdly, it has long been known that even in organisms with similarly strong selected codon usage bias often use different optimal codons. Why different synonymous codons can be optimal in different bacterial species

was still, largely, unclear. The process by which optimal codon usage could change was also unclear. The final part of this thesis is concerned with investigating changes in optimal codon usage across bacterial genomes as well as investigating the relationship between codon usage and the abundance of corresponding tRNA species.

Chapter 2:
Materials and Methods

2.1 Obtaining sequences

The *ACNUC* sequence retrieval software (Gouy *et al.*, 1985) was used to obtain bacterial genome sequences datasets from the *GenBank/EMBL/DDBJ* online databases. All coding sequences for each genome were extracted from *GenBank* using the genome accession number and the 't=CDS' attribute. In order to ensure that no copies of the 40 ribosomal protein and elongation factor genes (*rplA-F*, *rplI-T*, *rpsB-T*, EF-*Ts*, EF-*Tu* and EF-*G*), used as a highly expressed gene dataset, were missed due to errors in annotation, entire genome sequences were extracted and used to create *BLAST* (Altschul *et al.*, 1990) databases. Amino acid query sequences for these 40 proteins (from a closely related species) were then used as part of a *tBLASTn* against the genomic nucleotide *BLAST* database in order to find any of these genes that had escaped annotation. Another possible problem in extracting true coding sequences involved the GenBank misannotation of a gene's start codon. To check for this each of the 40 genes were separately aligned with those of closely related species, using ClustalW (Thompson *et al.*, 1994), to check for obvious errors in start codon position annotation.

In addition tRNA gene copy numbers were obtained from the *Genomic tRNA database* (http://lowelab.ucsc.edu/GtRNAdb/) whilst aligned 16S rRNA gene sequences were obtained from the *Ribosomal Database Project II release 9* (http://rdp.cme.msu.edu/) (Olsen *et al.*, 1991).

2.2 Analysis of intragenomic codon usage variation

2.2.1 Using CodonW to explore Codon Usage

The main tool used in this thesis to analyze codon usage was the *CodonW* package (Peden, 1999). This package allows sophisticated analyses of

codon usage and calculates values such as GC3s, RSCU, Nc, Fop, CAI, GRAVY and AROMO. These terms are defined in the following section.

2.2.1.1 Synonymous site composition statistics

Synonymously variable third positions refer to amino acids with synonymous codons such that a change in the base at the third codon position may not change the amino acid encoded (i.e. not Met or Trp). The two main calculations performed investigating base compositions at silent sites were GT3s and GC3s. The GC3s value is the fraction of codons, which are synonymous at the third codon position and have either a G or a C at that codon position. Similarly, the GT3s value is the fraction of codons, which are synonymous at the third codon position and have either a G or a T at that codon position. The GC3s and GT3s values can be calculated using the following equations:

$$GT3s = \frac{\left[\left(NNU + NNG\right) - \left(AUG + UGG + UAG\right)\right]}{\left[NNN - \left(AUG + UGG + stop\right)\right]}$$

$$GC3s = \frac{\left[\left(NNC + NNG\right) - \left(AUG + UGG + UAG\right)\right]}{\left[NNN - \left(AUG + UGG + stop\right)\right]}$$

Where NNU, NNG, NNC etc. refer to the total number of codons of that form.

2.2.1.2 Relative Synonymous Codon Usage

Relative Synonymous Codon Usage (RSCU) is calculated as the observed codon usage divided by the average codon usage for that amino acid (see equation). A value of 1.00 is obtained if all codons for a particular amino acid are used equally. RSCU removes the influence of amino acid composition that is present in raw codon usage data (Sharp & Li, 1986b).

$$\mathrm{RSCU} = \frac{n_i X_{ij}}{\sum_{J=1}^{n_i} X_{ij}}$$

Where X_{ij} is the frequency of the jth codons for the ith amino acid, encoded by n_i synonymous codons.

2.2.1.3 Effective Number of Codons (N_C)

In order to calculate the effective number of codons (N_C) Wright, 1990 the homozygosity for each amino acid is estimated from the squared codon frequencies:

$$F = \frac{\left(n\sum_{i=1}^{k} p_i^2 - 1 \right)}{(n-1)}$$

Where k = number of synonyms; *n* = total usage of k-fold synonymous amino-acid; *F* = homozygosity; p_i = frequency of usage of synonymous codon *i.*

The genetic code has five amino acid family types (non-synonymous, 2-fold, 3-fold, 4-fold and 6-fold synonymous amino acids). The N_C value is calculated as the arithmetic average of all non-zero homozygosity values for each of the amino acid family types.

$$N_C = 2 + \frac{9}{\overline{F}_2} + \frac{1}{\overline{F}_3} + \frac{5}{\overline{F}_4} + \frac{3}{\overline{F}_6}$$

Where F_i is the average homozygousity for the class with *i* synonymous codons

The effective number of codons provides a way to quantify how different the codon usage of a particular gene is from the equal use of synonymous codons. N_C is an estimate of the strength of general codon usage bias, and might be influenced by mutation biases and/or selection for particular codons.

The calculation of an N_C value for each gene in a genome shows how restricted synonymous codon usage is for that gene. However, this restriction of codon usage is sometimes interpreted as evidence for selection when the bias may be due to the effect of a mutational process instead. In order to account for the effect of G+C bias on this index one can use the equation below to calculate the expected value of N_C if codon bias is solely a function of GC3s.

$$EN_C = 2 + S + \left(\frac{29}{S^2 + (1-S)^2} \right)$$

Where S is the frequency of G+C (i.e. GC3s)

This equation can be overlaid onto a plot of N_C vs. GC3s and genes with codon choice constrained by a G+C mutation bias alone will lie on or just

below the line. The N_C value can be calculated from codon usage data alone and values of Nc range from 20 to 61 (Wright, 1990). A value of 20 is arrived at when only one synonymous codon is used for each amino acid and 61 when all synonymous codons are used equally.

2.2.1.4 Codon Adaptation Index (CAI)

The codon adaptation index (CAI) estimates the strength of selected codon usage bias within a gene (Sharp & Li, 1987a). In order to calculate CAI values first RSCU values (see section 2.2.1.2) should be calculated from very highly expressed genes of the organism in question. The relative adaptiveness of a codon is then given by the frequency of that codon compared to the frequency of the optimal codon for that amino acid (using this dataset of highly expressed genes).

$$. \omega_{ij} = \frac{RSCU_{ij}}{RSCU_{i\max}} = \frac{N_{ij}}{N_{i\max}}$$

Where ω_{ij} is the relative adaptiveness of the jth codon for the ith amino acid. $RSCU_{ij}$ is the RSCU value for the jth codon of the ith amino acid. $RSCU_{imax}$ is the RSCU value for the most frequently used codon from the highly expressed reference dataset for the ith amino acid. N_{ij} is the frequency of the use of the jth codon for the ith amino acid. N_{imax} is the most frequently used codon from the highly expressed reference dataset for the ith amino acid.

The CAI value for a gene is then given by the geometric mean of the relative adaptiveness values of each of the codons present in the gene.

$$\mathrm{CAI} = \exp\left(\frac{1}{L}\sum_{k=1}^{L}\ln\omega_k\right)$$

Where ω_k is the relative adaptiveness of the kth codon and L is the number of synonymous codons in the gene

A maximum CAI value of 1.0 is indicative of the use of only the most frequently used codons seen in the putatively highly expressed dataset.

2.2.1.5 Fop

The calculation of the frequency of optimal codons first requires the identification of optimal codons for the genome in question. Ikemura originally defined highly expressed genes based on tRNA abundances due to the correlation in *E. coli* between codon usage and tRNA content (Ikemura,

1981b). However, this can also be achieved by comparing codon usage in a dataset of highly expressed genes with that of the genome as a whole (see section 2.2.5) in order to identify codons used at significantly higher levels in highly expressed genes as compared to the genome as a whole. Once optimal codons have been defined F_{OP} is given by the frequency of optimal codons in a gene divided by the total number of codons in the gene.

$$\mathrm{F_{OP}} = \frac{N_{opt}}{N_{tot}}$$

Where N_{opt} = total frequency of predefined optimal codons in a gene, and N_{tot} = total frequency of synonymous codons in a gene

2.2.1.6 GRAVY

GRAVY is a measure of hydrophobicity of the hypothetical protein coded for by the gene (Kyte & Doolittle, 1982). It is calculated as the average of the hydrophobicity values of all of the amino acids present in the protein.

$$\mathrm{GRAVY} = \frac{1}{N}\sum_{i=1}^{N} k_i$$

Where N is the number of amino acids used in the hypothetical protein product and k_i is the hydrophobicity index of the ith amino acid

2.2.1.7 AROMO

AROMO is a measure of the aromaticity of the hypothetical protein coded for by the gene in question. It is calculated as the fraction of aromatic amino acids (Phe, Tyr, Trp) present in a protein.

$$\mathrm{AROMO} = \frac{1}{N}\sum_{i=1}^{N} v_i$$

Where N is the number of amino acids used in the hypothetical protein product and v_i is either 1 if an aromatic amino acid is being considered or 0 if not.

2.2.2 Correspondence Analysis

Correspondence analysis is a form of multivariate statistical analysis described by Greenacre (Greenacre, 1984), which allows a sophisticated way to explore complex datasets. Correspondence analysis is largely a graphical approach as opposed to a statistical one. This method is a technique used to visualize and explore complex datasets so that

correlations in the data can be examined. It is up to the user to infer what these correlations indicate and what is causing such a correlation.

The basic function of correspondence analysis is to reduce a multidimensional space into a lower dimensional subspace that best represents the variation among the data points. This is done by calculating an eigen vector that passes through or closest to the greatest number of points in the multidimensional space; this line is termed axis 1. The process is repeated with another eigen vector (axis 2) being calculated that is orthogonal (perpendicular) to axis 1. The process is then repeated until no further lines can be drawn (giving 41 axes for correspondence analysis on RSCU data).

Correspondence analysis was performed on RSCU data to overcome the effect of biases in amino acid composition. The analysis begins with a codon usage matrix that has dimensions X (number of genes) by Y (Codon usage values). As Met, Trp and stop codons are excluded Y is reduced to 59. However with RSCU values one loses one degree of freedom for each amino acid, because the values sum to the number of synonyms, thus there are 41 independent variables going into the analysis and therefore 41 axes come out. In practice, however, only the first three or four axes have been found to reflect interpretable biological variations in codon usage. Each gene has a coordinate on each of these new axes. Correspondence analysis also produces coordinates for codons and these two plots (genes on axes *a* and *b*, and codons on axes *a* and *b*) 'correspond', so that it is possible to visualize which codons are responsible for the spread of genes along axes.

2.2.3 Within-Block Correspondence Analysis

Within-block correspondence analysis was carried out using the *ade4* package (Thioulouse *et al.*, 1997) for the statistical environment *R* (http://www.r-project.org/). Performing a within-block correspondence analysis is essentially a two stage process. Firstly, a correspondence analysis is performed on a table of raw codon counts for each gene and then a within-block correspondence analysis is performed on the modified datatable produced by the correspondence analysis procedure.

Given a table of raw codon counts, with rows representing each of the 59 codons (64 possible codons minus 3 stop codons, Met and Trp) and columns representing each of the genes in the genome, the table's rows and columns are weighted appropriately, in the *dudi.coa*() procedure, using the equation below.

$$FinalValue = \frac{k_{(ij)} \times k}{k_{(i)} \times k_{(j)}} - 1$$

where $k_{(i,j)}$= each cell, k = the grand total of rows and colums (raw counts) , $k_{(i)}$ = the row totals (raw counts), $k_{(j)}$ = the column totals (raw counts)

In this resultant table each of the values is then multiplied by the original row weightings ($k_{(ij)}/k$) to give a new table. The next step is to call the within-block correspondence analysis procedure in the *ade4* package (*dudi.within()*) which uses this modified datatable along with a vector supplied to the procedure to distinguish which codons should be grouped as synonymous in the analysis.

Synonymous codon rows have their individual values for each synonymous codon summed (to give totals for amino acids) and then this value divided by the summed original row weights for each group of synonymous codons (i.e. summed codon row weights give total weightings for each amino acid) to give a final weighted table with 18 rows (20 amino acids minus methionine and tryptophan). This final table is then used to perform a correspondence analysis as normal.

2.2.4 GC Skew

In bacterial genomes the leading strand is often more G+T rich than the lagging strand and so by observing change in G-C skew across the genome is often possible to predict the positions of the origin and/or terminus of replication (Lobry, 1996; McLean *et al.*, 1998; Picardeau *et al.*, 2000). The G-C skew at synonymously variable third positions was calculated for each gene using the equation below and then a plot of G-C skew across the genome was produced using a moving average with a 50 gene window.

$$\mathrm{GC}_{\mathrm{SKEW}} = \frac{(G-C)}{(G+C)}$$

2.2.5 Identification of optimal codons

Putative optimal codons were identified by comparing codon usage in the 40 highly expressed gene dataset (*rplA-F*, *rplI-T*, *rpsB-T*, EF-*Ts*, EF-*Tu* and EF-*G*) against the genome as a whole. Codons used at higher frequencies in the highly expressed dataset as compared with the genome as a whole were then put forward as potential optimal codons and tests were carried out to assess whether such codons were significantly over used in highly expressed genes.

To do this, χ values were calculated comparing codon usage in the putatively highly expressed genes with the codon usage in the total dataset. However due to the multiplicity of the number of tests performed probability values need to be adjusted. The standard approach is the Bonferroni correction, but this has been judged to be too stringent. Rice suggested a method where the Bonferroni correction is applied sequentially, rather than simultaneously to all χ^2 values (Rice, 1989). This has been used before, in the context of codon usage analysis (Grocock & Sharp, 2002; Henry & Sharp, 2007).

2.3 Analysis of Intergenomic Codon Usage Variation

2.3.1 The strength of selected codon usage bias

Sharp and co-workers (Sharp *et al.* 2005) devised a method to measure the strength of selected codon usage bias within a genome. Sharp's logic behind the creation of this strength of selected codon usage bias statistic was as follows:

Using Bulmer as a basis (Bulmer, 1991), one can consider the case of an amino acid encoded by two synonyms, C_1 and C_2. The mutation rate from C_1 to C_2 is u, and the mutation rate from C_2 to C_1 is v. The selective difference between the two codons is s, the fitness of optimal codon C_1 is 1, while that of C_2 is (1 - s). Under the combined effects of mutation,

selection and random genetic drift, the equilibrium frequency (*P)* of C_1 in a gene or set of genes, is given by:

1)

$$P = \frac{e^s V}{\left(e^s V + U\right)}$$ **where *S*=2*N*$_e$*s*, *U*=2*N*$_e$*u* and *V*=2*N*$_e$*v***

In genes where selection is strong enough to influence codon usage, the frequency of codons is determined by both the pattern of mutation and the strength of selection. The magnitude of *S* can be estimated from

2)

$$S = \ln\left[\frac{\left(P_H \times k\right)}{\left(1 - P_H\right)}\right]$$ **where *k* = *U/V* and P_H is the frequency of the C_1 codon in the highly expressed gene dataset**

In genes where selection is so weak as to be ineffective, the frequency of the codons is determined by the pattern of mutation between them:

3)

$$P_L = \frac{V}{\left(V + U\right)}$$ **where P_L is the frequency of the C_1 codon in the dataset as a whole (taken to be under low selection)**

This allows the estimation of k using:

4)

$$k = \frac{\left(1 - P_L\right)}{P_L}$$

For use in equation 2 above

This methodology was applied to codons for four two-fold degenerate amino acids where synonymous codons take the form WWY (i.e. Phe, Tyr, Ile and

Asn). Across all bacterial species the WWC codon is preferred to the WWU codon as only one tRNA species is present in the genome to decode both codons. The first anticodon position of this tRNA is a guanine and so it pairs exactly with the WWC codon whilst pairing with WWU through wobble, assuming no base modifications occur. This means that WWC is always better recognized, even if (due to the absence of effective selection) it is not always seen to be preferred, and hence is translationally optimal as it promotes more accurate and efficient translation. It is worth noting that isoleucine is, in fact, three fold degenerate but its third codon (AUA) is generally rarely used and so isoleucine can be considered as effectively two fold degenerate.

The strength of selection, 'S', was measured by comparing genes that one would expect to be influenced heavily by selection (i.e. ribosomal protein genes and elongation factors, in this case *rplA-F*, *rplI-T*, *rpsB-T*, EF-*Ts*, EF-*Tu* and EF-*G*) and comparing them to the genome as a whole, which should be broadly under low selection, as the number of genes expressed at high levels is a very small fraction of the genome as a whole. For each of the four amino acids 'S' was calculated using the equations above, with the WWC codon as the C_1 codon and WWU as the C_2, codon in each case. Individual 'S' values for each codon were then weighted by abundance in the highly expressed dataset and summed to give a final value S_{WWY}.

2.4 Methods for phylogeny construction

2.4.1 MrBayes

The phylogeny construction program *MrBayes* (Ronquist & Huelsenbeck, 2003) was used to construct bacterial phylogenies by Bayesian methods. In a Bayesian analysis a phylogeny is inferred based on the posterior probabilities of the phylogenetic trees which can be expressed using Bayes's Rule as follows:

$$f(\tau,\nu,\theta \mid X) = \frac{f(\tau,\nu,\theta) f(X \mid \tau,\nu,\theta)}{f(X)}$$

Where $f(\tau,\nu,\theta)$ is the prior distribution (specifying the prior probability of the different parameter values), $f(X \mid \tau,\nu,\theta)$ is the likelihood function (describing the probability of different parameter values), $f(X)$ is the total probability of the data summed and integrated over the parameter space and $f(\tau,\nu,\theta \mid X)$ is the posterior distribution.

It is often not possible to calculate posterior probabilities analytically and so a Markov chain Monte Carlo (MCMC) method is used to obtain samples from it. The MCMC method works by altering tree parameters such as tree topology (τ), branch lengths (ν) and substitution parameters (θ) using a stochastic mechanism. Once a parameter change has been made the change is either kept or discarded based on the change in likelihood. *MrBayes* implements a variety of stochastic models for nucleotide, protein, restriction site and morphological data. Rate variation across sites can also be accommodated using a standard gamma distribution. Individual parameter values are given in the materials and methods sections of individual results chapters where bacterial phylogenies are presented.

As discussed, confidence in a *MrBayes* constructed phylogeny is assessed using posterior probabilities based on the frequencies with which parameter values are observed. *MrBayes* was run on the Nottingham University HPC cluster using 8 individual nodes performing 2 distinct runs and 4 chains for each run. This was possible due to the Metropolis-coupled MCMC method employed by *MrBayes* which uses several chains which can be 'heated'. Heating is defined as the proportional, exponential increase in the posterior probability of a step (Ronquist & Huelsenbeck, 2003). Such a method allows individual chains to escape local valleys on the likelihood surface where a step-wise change would not. In *MrBayes* these chains are able to communicate with each other and chain states can be swapped based on differences in likelihood.

2.4.2 Assessing the progress of *MrBayes*

Tracer is an application created for use with the program *Beast*, a program for Bayesian MCMC analysis of molecular sequences created by Andrew Rambaut and Alexei Drummond and is freely available from the Oxford University evolutionary biology group website (http://evolve.zoo.ox.ac.uk/software.html). *Tracer* is also suitable for analyzing the progress of a *MrBayes* run to assess whether a run is nearing completion. Likelihood values are displayed at intervals during the run: when the likelihood trace reaches and maintains a maximum level it is likely that an optimal tree has been arrived upon. *MrBayes* performs two runs

simultaneously and so if the two independent runs also converge upon the same position then one can be even more confident.

2.5 Tools for exploring the data

The output of the *CodonW* package (Peden, 1999) is a series of flat text files. To explore the data and find trends within it, the output from *CodonW* was imported into *Microsoft Excel*. *MS Excel* is not always efficient at importing data from multiple large text files, especially when more sophisticated levels of automation such as the production of graphs are required. For this reason the output of the *CodonW* package was manipulated using Perl and *VBA* scripts.

2.5.1 *Perl* and *VBA* scripts

Perl scripts were written to perform a number of basic functions throughout the course of this thesis. They were primarily written to handle the output of the *CodonW* (Peden, 1999) package and format the data so that it could be passed on to *VBA* scripts and be imported in *Excel*. As an added feature simple calculations were performed on the data such as the calculation of Pearson correlations and genomic G+C content, before exporting to *Excel*. The output of the *Perl* scripts produced a few large text files that were then read straight into *Excel*, formatted, and simple plots such as the Nc plot were produced automatically using VBA scripts. Scripts were also used to perform a variety of simple tasks, such as to calculate total codon usage in a genome by summing the codon usage of each individual gene.

2.5.2 The *R* statistical environment

R is a programming language and software environment for statistical computing and graphics. It is a GNU project related to the *S language*, an environment which was developed at Bell Laboratories by John Chambers and colleagues. The *R* environment was used for analysis, such as the within block correspondence analysis using the *ade4* package (Thioulouse *et al.*, 1997) as well as to construct many of the plots used in this thesis.

Chapter 3: Codon usage variation in the genome of Bdellovibrio bacteriovorus

3.1 Introduction

This chapter aims to investigate factors affecting intragenomic codon usage patterns in *Bdellovibrio bacteriovorus*. The techniques used to do this are well established and have been used to look at codon usage patterns for many other bacterial genomes.

3.1.1 Bdellovibrio bacteriovorus

The Delta Proteobacterium *Bdellovibrio bacteriovorus* lives in a wide variety of environments. Its name can be translated as 'curved leach' and this is a rather appropriate description, due to its distinctive shape and lifestyle. *B. bacteriovorus* is a highly motile bacterium that preys on other Gram-negative bacteria such as *Escherichia coli* (Stolp & Starr, 1963). *Bdellovibrio* were originally discovered by Stolp and Starr in soil samples and since then other isolates have been found from marine sediments, rivers and plant rhizospheres, as well as a variety of other habitats including the intestinal tract of mammals (Rendulic *et al.*, 2004). Therefore it can be seen that these organisms are extremely abundant in nature. Interest in *Bdellovibrio* has been great due to its ability to kill other pathogenic bacteria whilst being unable to infect mammalian cells (Lenz & Hespell, 1978); this has led to it being dubbed a 'living antibiotic'.

3.1.1.1 Life Cycle

The life cycle of *Bdellovibrio* has two major phases, the attack phase (Figure 3-1 labels I-IV) and the growth phase (Figure 3-1 labels V-VIII). Whilst in the attack phase the bacterium has a flagellum and is highly motile. The bacterium locates prey via chemosensors (I) and violently collides with its prey. A reversible attachment is initially formed followed by an irreversible attachment (II) after a short 'recognition period' (Rendulic *et al.*, 2004). Following irreversible attachment *Bdellovibrio* breaches the

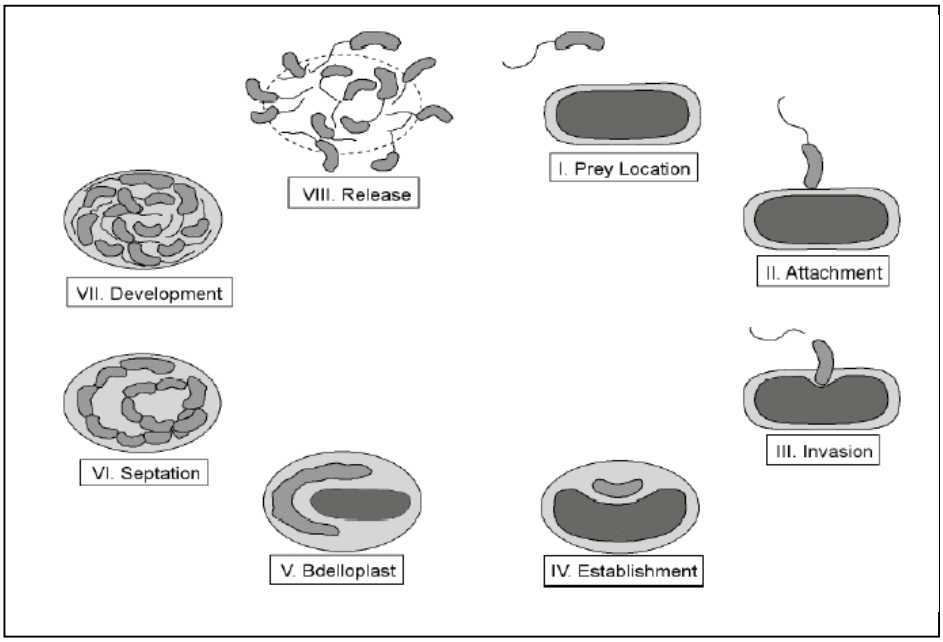

Figure 3-1 The Life Cycle of Bdellovibrio.
Figure showing the eight key stages in the lifecycle of *Bdellovibrio.* Taken from Rendulic et al, 2004

outer membrane and kills its prey by halting its respiration and growth (III). After breaching the outer membrane the host loses its flagellum and enters 'growth phase' (Nunez *et al.*, 2003; Rendulic *et al.*, 2004). The bacterium then takes up residence between the inner and outer membranes in the prey's periplasm (III-IV). The predator begins to alter the prey's exterior membrane and peptidoglycan layer but crucially does not destroy it. It is thought that these large scale outer membrane modifications take place to preclude the entry of any other invading bacterium. It has been reported that multiple invasions are possible, but this is thought to be due to two *Bdellovibrios* invading distal ends of a prey bacterium almost simultaneously (Nunez *et al.*, 2003), this would not give time for large scale membrane modifications to take place and so does not disprove the initial theory. The process of modification causes the invaded cell to swell and form the characteristic shape of the Bdelloblast (V).

As well as modifications to the prey's outer membrane, the inner membrane is additionally modified. The cytoplasmic membrane is altered so that the predator can insert degradative enzymes and extract degraded cell material. The predator degrades the prey's DNA and RNA into their nucleotide constituents so that they can be used by the predator to synthesize its own nucleic acids (Beck *et al.*, 2004). In addition to nucleotides, fatty acids are also thought to be taken up by *Bdellovibrio*. The bacterium is only thought to be able to synthesize 11 of the amino acids needed for protein synthesis itself and so is thought to extract amino acids, or at least amino acid precursors, from the host cell. This means that *Bdellovibrio bacteriovorus* can only synthesize proteins if it has access to a host. Outer membrane proteins and lipopolysaccharides of the prey are not thought to be reutilized by the predator. It is thought that integration of outer membrane proteins belonging to the prey would most probably affect the *Bdellovibrio* detrimentally. The outer membrane of the prey cell is maintained by the predator, although its structure may be significantly altered. It is thought that the maintenance of this membrane prevents the diffusion of nutrients away from the bdelloblast and therefore is beneficial to the predator.

The *Bdellovibrio* cell forms a long filament as it grows and eventually septates (VI). The progeny continue to develop (VII) into cells complete with flagella from within the prey protoplast. At this stage *B. bacteriovorus* produces hydrolytic enzymes that dissolve the peptidoglycan layer and the outer membrane (VIII) so that the prey can escape (Rendulic *et al.*, 2004). As many as 15 bacteria can be released and they then search out new host cells and so the life cycle continues. It may be expected that such an organism with a rapid generation time of around 30 minutes would have codon usage patterns heavily influenced by selected codon usage bias.

3.1.1.2 Strain under study

The strain of *Bdellovibrio bacteriovorus* that is under study here is the host dependant strain HD100 (*GenBank* accession number: BX842601). This strain, as its name suggests, requires prey to be able to complete its life cycle and reproduce. The genome of this bacterium is one large chromosome of 3,782,950 base pairs, has 2 rRNA genes, 36 tRNA genes and encodes 3584 proteins. This is of a similar size to other saprophytic bacteria despite the predatory nature of *B. bacteriovorus*. This suggests that predation is a lifestyle choice with *Bdellovibrio* having evolved from saprophytic ancestors by the acquisition of predatory genes while retaining the ability to slowly grow in the prey-independent state.

3.1.2 Aims of this study

The primary aim of this study was to learn more about *Bdellovibrio* and the genes that enable it to have such an unusual life cycle. This bacterium was also chosen as its genome had just been completely sequenced at the time I was beginning my PhD work. A colleague in the School of Biology, Professor Liz Sockett, was involved in with the sequencing of the genome and *Bdellovibrio* is one of the main focuses of her research. Therefore, it made sense to collaborate with the Sockett lab and try to find potentially important genes using computational means that could be further investigated experimentally by the Sockett lab.

An additional aim was to explicitly compare two forms of correspondence analysis. The Sharp lab has published many papers using correspondence

analysis (CA) on relative synonymous codon usage (RSCU) data to look at patterns of codon usage in bacterial genomes (Recently in: Grocock & Sharp, 2001 ; Grocock & Sharp, 2002; Henry & Sharp 2007). However it has been argued that such a method is flawed as correspondence analysis should be performed on raw counts only and not normalized data such as RSCU normalized data. An alternative method of CA was suggested which, it was argued, was better suited to the analysis of codon usage patterns (Perrière & Thioulouse, 2002). The analysis of factors affecting codon usage in *Bdellovibrio bacteriovorus* gave an opportunity to compare these two methods to discover whether CA on RSCU data necessarily gives bad results or whether it depends on how carefully those results are examined and interpreted.

3.2 Specific Materials and Methods

All methods described in this chapter were done in accordance with those described in chapter two entitled Materials and Methods. Any deviations or additions to those methods specific to this chapter are described here.

3.2.1 Orthologue detection details

Two programs were used to find orthologous gene pairs between *B. bacteriovorus* and *E. coli*. One method used the *inparanoid* program (Remm *et al.*, 2001), the other the HOGENOM database (Jan 28th 2005) (Dufayard *et al.*, 2005). As input *inparanoid* takes two files each containing all protein sequences encoded by the genes in a particular genome. The program works by first performing an all-against-all *blast* (Altschul *et al.*, 1990) search between two genomes and selects matches with a cut-off of 50 bits with an overlap of more than 50% of the gene. Bit scores are a useful way to compare different alignments as this score accounts for the type of scoring system used; the bit score is calculated from the raw alignment score but normalized with the statistical variables that define a given scoring system. The authors of the program state that a score of 50 bits was decided upon for empirical reasons as this value generally removes the majority of insignificant hits, thus reducing the CPU load when clustering is then performed. They also state that the overlap cut-off should be set to 50% overlap to avoid the inclusion of short, domain level

matches. Finally, such results are clustered and the two-way best hits are given confidence values. This method predicted 991 orthologous gene pair matches between the two genomes.

The *HOGENOM* database was accessed through the *FamFetch* program (http://pbil.univ-lyon1.fr/software/famfetch.html) and the database was filtered for gene families present in both *B. bacteriovorus* and *E. coli*. This initial approach gave 713 potential gene family matches. The *HOGENOM* flat database files were also obtained for the two genomes so that genes in each family could be associated with their unique gene identifier names.

Genes found by both methods were combined and duplicate matches found by both techniques were removed. Only matches assigned a 100% confidence value by *inparanoid* were kept and any gene families from *HOGENOM* that contained more than one gene from each genome were discarded. This process gave 1061 predicted orthologous gene pairs.

3.2.2 Comparing gene orthologue pairs for differences in codon usage patterns

To look for gene pairs with different putative expression patterns the F_{OP} values for each of the homologous genes found between *E. coli* and *B. bacteriovorus* were plotted. A model II regression line (see section 3.2.3) was fitted through the data. To look for genes with high F_{OP} values in *Bdellovibrio* but not in *E. coli* it had to be decided how far away the gene need to lie away from the regression line to be significant.

One important thing to take into account was gene length as a shorter gene has a higher chance of having a larger deviation just by chance, due to the way F_{OP} is calculated. The expected standard error (SE) for a F_{OP} value can be derived from the binomial:

$$SE = \sqrt{\frac{P(1-P)}{L}}$$

Where P is the fraction of optimal codons in a gene (i.e., $p=F_{OP}$), and L is the number of codons for amino acids included in the calculation (i.e., excluding Met, Trp, stop and any codons for AA that do not have optimal codons assigned).

It is then simple to look for genes that lie more than 2 SE above (or below) the regression line, with gene length accounted for. However, this takes no account of the error (of the same source) attached to the *E. coli* F_{OP} value. To take account of the contribution of the error in the *E. coli* F_{OP} value it was multiplied by the regression coefficient and added to the *Bdellovibrio* error (Brookfield & Sharp, personal communication).

$$SE = \sqrt{\frac{P_B(1-P_B)}{L_B} + b^2\left[\frac{P_E(1-P_E)}{L_E}\right]}$$

Where p_B and p_E are the F_{OP} values for the *Bdellovibrio* and *E.coli* orthologues, L_B and L_E are their lengths in (relevant) codons.

Since the variance of the *Bdellovibrio* and *E. coli* F_{OP} values is likely to be similar, if the slope of the regression line is about 1.0 , the new cut off (2 SE) will be approximately 1.4 times that not taking account of the *E. coli* error (about 0.14 for L=100, about 0.07 for L=400).

3.2.3 Model II regression

Model II regression is used when the two variables in the regression equation are uncontrolled by the researcher. If this is true the data contained in both axes is subject to error and if Model I regression were used an underestimate of the slope of the linear relationship between the variables. The form of Model II that was used in this thesis is known as Reduced Major Axis regression. It was calculated using an Add-in for Excel created by M. Sawada of the University of Ottawa (http://www.lpc.uottawa.ca/data/scripts/).

3.3 Results

3.3.1 Overview of Codon Usage bias in *Bdellovibrio bacteriovorus*

The *B. bacteriovorus* genome is comprised of one large chromosome that encodes 3854 proteins. Short genes coding for proteins less than 50 amino acids in length were removed from this analysis to leave a genome containing 3512 genes. When these remaining genes are considered the genome is slightly G+C rich with a genomic G+C content of 0.51 and a G+C

content at synonymously variable third positions, GC3s, of 0.56. There are some sections of the genome that show an abnormal G+C content, these sections are unusually A+T rich and code for LPS synthesis proteins (Bd1678-Bd1699), ribosomal genes and elongation factor genes (primary A+T rich cluster: Bd2949-Bd2994), restriction modification genes (Bd3691-3697) and another A+T rich region containing several hypothetical protein genes (Bd2672-Bd2682). These sections can easily been seen when the GC3s of each gene in the genome is plotted and a moving average trend line, with a 50 gene window, overlaid (Figure 3-2). When these LPS, RMS, ribosomal genes and hypothetical gene cluster are plotted onto the graph it can be seen that they correlate strongly with the dips in GC3s shown by the trend line, which are due to the extremely low GC3s values. Some of the genes in these clusters are below 0.30 and the majority of the genes have a GC3s of less of 0.45.

The plot of GC skew against genome number shows that the genome is split into two approximately equal parts (Figure 3-3). This switch in GC skew is due to a common feature of bacterial genomes where genes on the leading strand are more G+T rich than those on the lagging strand (Rocha *et al.*, 1999). This feature can be used to find the origin and terminus of replication using such a GC skew plot. The origin of replication is at position 0 of the sequence as indicated by the first gene being *dnaA* (Bd0001). From the G+C skew plot it was inferred that the terminus of replication was located approximately 1850 genes from the origin (Bd2027). Once the location of the origin and terminus of replication were assumed genes were assigned to the leading and lagging strands. In order to ensure accuracy when plotting leading and lagging strand genes, 50 genes either side of the putative terminus and origin were not included in the plot as one could not be sure of the exact location of the terminus and origin of replication. The average weighted GT3s value, treating all codons as one concatenated super-gene and excluding methionine, tryptophan and the stop codons, for the leading strand was 0.553 for the leading strand and 0.471 for the lagging strand, illustrating the extent of the strand bias.

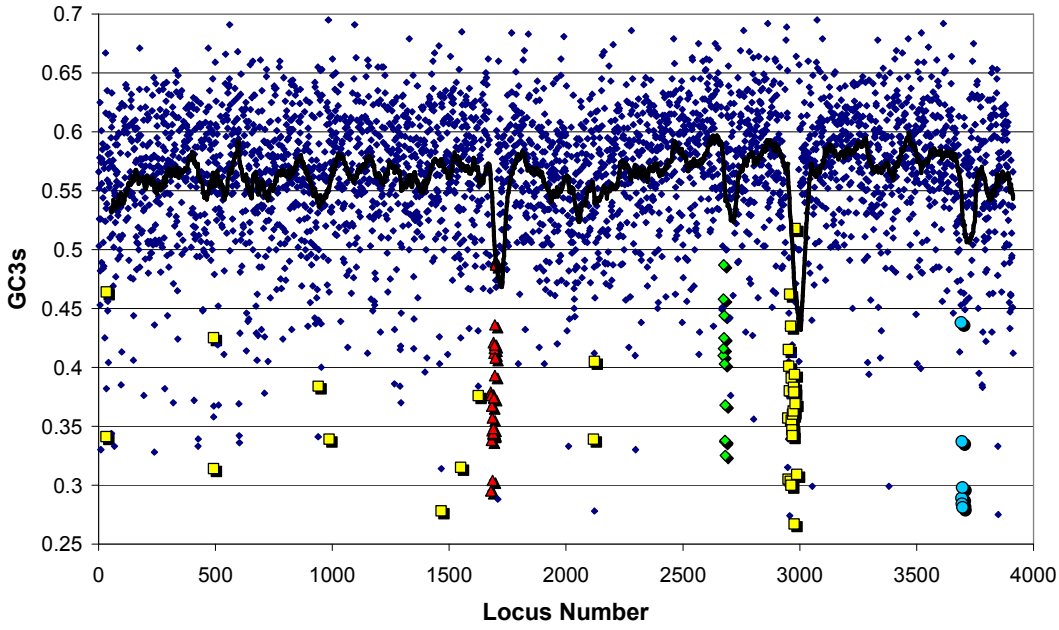

Figure 3-2 Plot of GC3s content across the *B. bacteriovorus* genome.
The putatively highly expressed gene group (*rplA-F*, *rplI-T*, *rpsB-T*, EF-*Ts*, EF-*Tu* and EF-*G)* are marked in yellow, LPS genes in red, RMS genes in blue and the final A+T rich region genes in green.

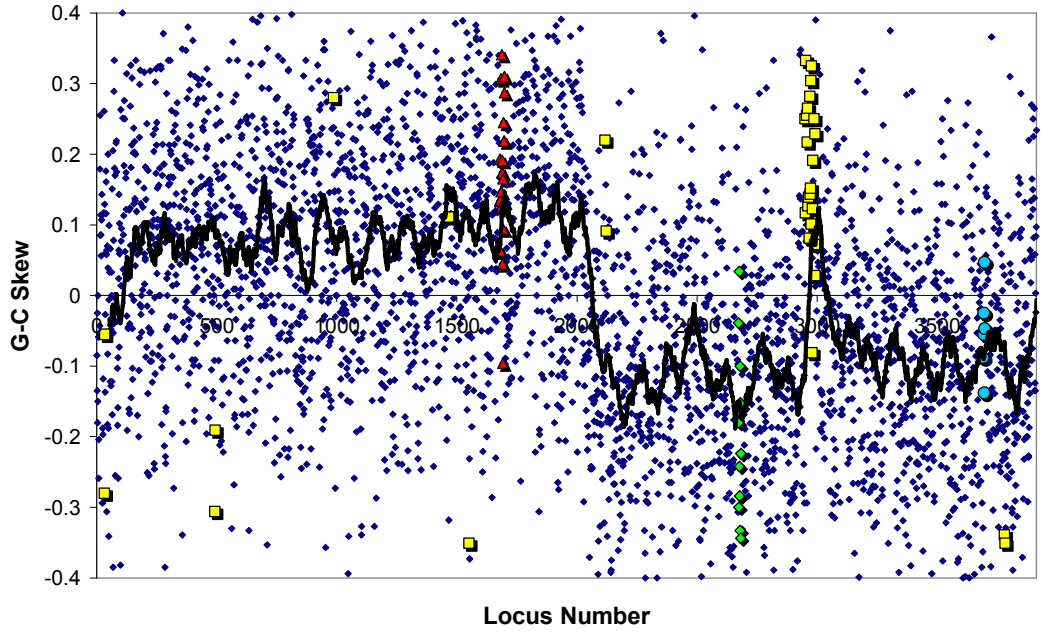

Figure 3-3 Plot of G-C Skew across the *B. bacteriovorus* genome.
The putatively highly expressed gene group (*rplA-F*, *rplI-T*, *rpsB-T*, EF-*Ts*, EF-*Tu* and EF-*G)* are marked in yellow, LPS genes in red, RMS genes in blue and the final A+T rich region genes in green. G-C Skew for strand '1' was calculated for each gene using G3s and C3s values and multiplying by -1 if the gene was on strand '2'.

In addition to the switch in G+C skew as a result of the terminus of replication an additional peak can be seen on the plot that coincides with the primary ribosomal gene cluster. The other A+T rich regions identified do not cause such a feature on the plot which suggests that the reasons for the A+T rich ribosomal genes have a different origin to those of the LPS and RMS gene clusters. The features identified using these plots will be discussed further later in the chapter.

3.3.2 Initial analysis of codon usage bias

As a preliminary treatment of the codon usage data a plot of N_C vs. GC3s was used to look for and examine codon usage heterogeneity in the *B. bacteriovorus* genome.

The N_C plot (Wright, 1990) was devised to look at codon usage heterogeneity and here one can clearly see differences in codon usage patterns within the genome (Figure 3-4). The plot shows that the majority of genes in the genome cluster in one main cloud centering on 0.5 to 0.6 GC3s with N_C values ranging from around 40 to 55. These N_C values are quite low and below the expected curve (the curve represents the expected N_C value based on GC3s alone) indicating that the many of the genes in the genome are subject to restricted codon usage. In addition to the main cluster there are other more widely distributed genes. Most of these genes with GC3s values that are different from the main cluster appear to be the A+T rich genes (red squares) and the ribosomal genes (yellow squares).

These ribosomal protein genes appear to be less G+C rich (around 0.3-0.4 GC3s) and also show more restricted codon usage than may be expected for their GC3s content, as indicated by the curve, with some having N_C values as low as 30. In addition there are genes with values even lower than these genes. However upon inspection many of these genes encode only hypothetical proteins and so may not actually be real genes. One gene that does stand out, however, is the gene coding for ATP synthase subunit C. Using a restricted subset of codons is often an indication of translational selection in the genome, where highly expressed genes, such as ribosomal

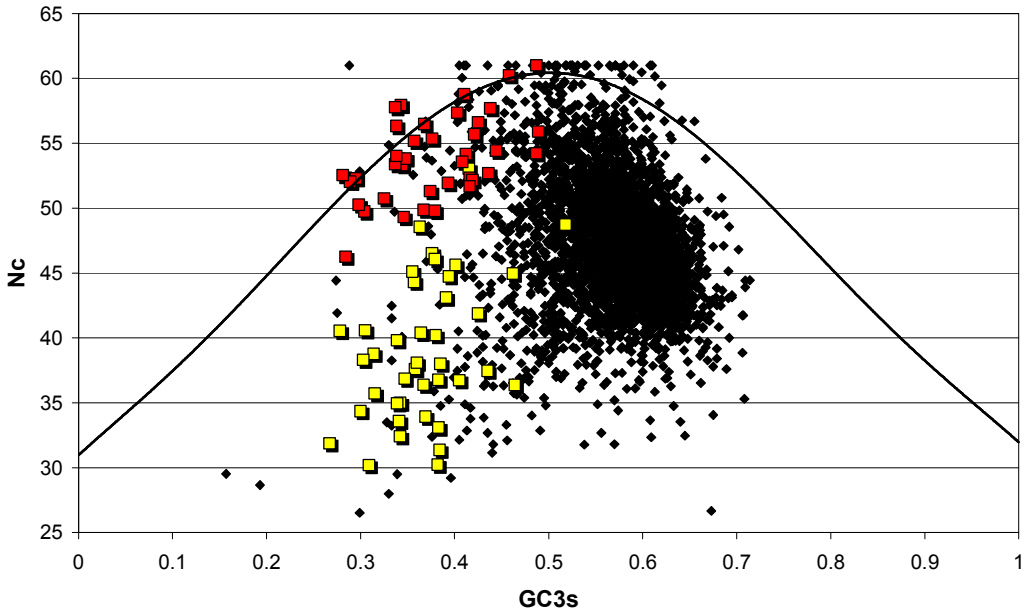

Figure 3-4 A plot of effective number of codons against GC3s for all the genes in the *B. bacteriovorus* genome
The N_C plot shows that many highly expressed genes (*rplA-F*, *rplI-T*, *rpsB-T*, EF-*Ts*, EF-*Tu* and EF-*G)*, shown in yellow, use a reduced set of codons. Previously identified A+T rich genes are marked in red.

protein genes, use a subset of codons that correspond to the most abundant tRNA species in the cell to ensure their accurate and/or efficient translation. In addition to these ribosomal genes and elongation factors genes (as well as additional ribosomal genes not included in the 40 highly expressed genes dataset), genes coding for ATP synthase subunits (ATP synthase subunits B and C with N_C values of 33.3 and 28.0) and chapronins (e.g. *groEL* and *groES* genes with N_C values of 32.9 and 34.8) were found. All these genes are potential candidates for the influence of selection due to their involvement in protein and energy production within the organism. In addition to this there are many genes with extremely high N_C values, some using the maximum 61 codons. Upon inspection this cluster is made up of genes encoding hypothetical proteins and as such cannot be guaranteed to be a genuine feature of the genome.

Although this plot indicated the possible presence of translational selection in the genome it is always sensible to carryout a full analysis of the data with multivariate statistical analyses.

3.3.3 Analysing patterns of codon usage using multivariate statistical analysis

In order to provide a thorough examination of factors influencing codon usage patterns in the *B. bacteriovorus* genome multivariate statistical analyses were performed. The program used for most of the codon usage analysis work presented here was the *CodonW* package (Peden, 1999). This program allows the user to perform a correspondence analysis (Greenacre, 1984) on RSCU data (see Chapter 2: Materials and Methods). Multivariate statistical analyses such as correspondence analysis (CA) are particularly well adapted to the multi-dimensional nature of the data and such a method is commonly used when analyzing codon usage biases in microbial genomes. However, it has been suggested that doing such an analysis on normalized data such as RSCU values is a misuse of correspondence analysis (Perrière & Thioulouse, 2002). Relative synonymous codon usage values are normally used instead of raw codon counts to avoid biases that are linked to amino acid composition, which may mask effects linked directly with synonymous codon usage. It has

been argued that correspondence analysis was originally designed to be performed on raw data counts and that using modified values can severely affect results. For this reason the analysis carried out here uses two techniques to analyze codon usage, one that uses normalized RSCU data and one that does not. The correspondence analysis carried out on the raw data uses a method known as within-block correspondence analysis (Thioulouse *et al.*, 1997) and is an alternative suggested by those who criticise the use of CA on RSCU data. It has been proposed that within-block correspondence analysis is able to remove the effects of amino acid bias without introducing unjustified statistical weights on data resulting in axes being generated that are primarily caused by differences in codon usage for rare amino acids such as cysteine. Such a method works by grouping synonymous codons together and weighting their influence depending on the amino acid abundance and is described in detail in the materials and methods section.

3.3.3.1 Codon usage analysis using correspondence analysis on RSCU data

In order to compare these two methods of correspondence analysis the *B. bacteriovorus* codon usage data was subject to both methods of multivariate statistical analysis. The CodonW package (Peden, 1999) was used to implement the correspondence analysis on RSCU data as described in the materials and methods chapter.

3.3.3.1.1 Axis 1 indicates translational selection

The plot of N_C vs. GC3s indicated that genes showing restricted codon usage were present in the genome (Figure 3-4). These genes included many ribosomal genes and some elongation factors, such as EF-Tu, which are known to be expressed at high levels within other bacterial genomes such as *E. coli*. This plot gave an indication that the restricted codon usage bias was a result of translational selection. In order to investigate further a correspondence analysis was carried out.

The first axis picked out by correspondence analysis on the RSCU data described 13.4% of the variation (Table 3-1). No extremely strong

correlations were present from the table between genomic features, such as base composition, and axis one. However, some positive correlation between axis one and GC3s (0.457), G3s (0.464) and N_C (0.445) was noticeable. The correlation between axis one and N_C suggests that the genes pulled to the left (negative) of axis one exhibit restricted codon usage bias.

This axis shows that the set of putatively highly expressed genes (*rplA-F*, *rplI-T*, *rpsB-T*, EF-*Ts*, EF-*Tu* and EF-*G*) known to be highly expressed in other genomes (Sharp *et al.*, 2005) are all located to one side of axis one (Figure 3-5). When the codons responsible for this axis were examined a broad range of codons were found to be responsible for this axis; with eight U-ending codons, five A-ending codons, one G-ending codon and six C-ending codons in the top 20 codons responsible for the pull of the highly expressed genes to the left on axis one; with GUA, UCU and CCA being the top three codons. Although the majority of these codons are A+U ending there were also G+C-ending codons responsible for this axis. It, therefore, does not appear that one particular base compositional bias is responsible for this axis. Instead, the cause of this axis is due to potentially highly expressed genes and the codons causing this axis were good candidate optimal codons as will be discussed later in this chapter. It is therefore likely that the main factor influencing codon usage in *B. bacteriovorus* is translational selection.

3.3.3.1.2 Axis 2 correlated with horizontally transferred genes

To investigate the trends responsible for the second axis the main codons involved in the axis were again examined. One end of axis two has mainly G+C ending codons (18 of the top 20 take the form NNG or NNC) whilst at the opposite end of the axis A+U ending codons are mainly found (17 of the bottom 20 take the form NNA or NNU). This initial investigation indicated that unusual base composition seemed to be the cause of this axis.

The second axis in the CA explains 8.1% of the total variation and a Pearson correlation between GC3s and axis two gave a very strong correlation of 0.68. Previous examination of the genome showed that certain areas of the genome contained genes that were particularly A+T

	Axis1	*Axis2*	*Axis3*	*Axis4*
Variation	*13.41%*	*8.05%*	*6.54%*	*3.81%*
GC3s	0.4568	**-0.6834**	-0.0257	-0.0220
GC	0.3460	**-0.5016**	-0.1140	0.0114
Gravy	0.1017	0.0861	-0.1352	0.0089
Aromo	0.1946	-0.0489	-0.0255	-0.0027
Nc	0.4448	0.4882	-0.0534	-0.0271
T3s	-0.3553	**0.5106**	-0.4074	0.0996
C3s	0.0521	**-0.5866**	**0.5470**	-0.0244
A3s	-0.2638	0.4177	0.4827	-0.0776
G3s	0.4638	-0.1528	**-0.6106**	0.0006
Y3s	-0.2907	-0.2052	0.2599	0.0698
K3s	0.1274	0.2407	**-0.7973**	0.0724

Table 3-1 Pearson correlations for correspondence analysis on RSCU data

Pearson correlations between each of the four major axes, created by the correspondence analysis on RSCU data, and various genomic feature are outlined in the table above. Figures in bold indicate those features that show a correlation of (absolute value) greater than 0.5

	Axis1	*Axis2*	*Axis3*	*Axis4*
Variation	*17.82%*	*10.39%*	*7.96%*	*2.91%*
GC3s	-0.4043	**0.7799**	0.2277	0.0003
GC	-0.2984	**0.5503**	0.2393	-0.0671
Gravy	-0.0966	-0.0974	0.1586	0.0136
Aromo	-0.1948	0.0564	0.0654	0.0207
Nc	-0.4950	-0.4193	0.0211	0.1294
T3s	0.3084	**-0.6804**	0.2818	-0.1478
C3s	0.0057	**0.7240**	-0.4751	0.0061
A3s	0.2402	-0.3696	**-0.6338**	0.1615
G3s	-0.4657	0.1164	**0.7638**	-0.0061
Y3s	0.3143	0.2033	-0.2975	-0.1399
K3s	-0.1627	-0.3929	**0.8335**	-0.1114

Table 3-2 Pearson correlations for within-block correspondence analysis on raw codon usage data

Pearson correlations between each of the four major axes, created by the within block correspondence analysis on raw codon usage data, and various genomic feature are out lined in the table above. Figures in bold indicate those features that show a correlation of (absolute value) greater than 0.5

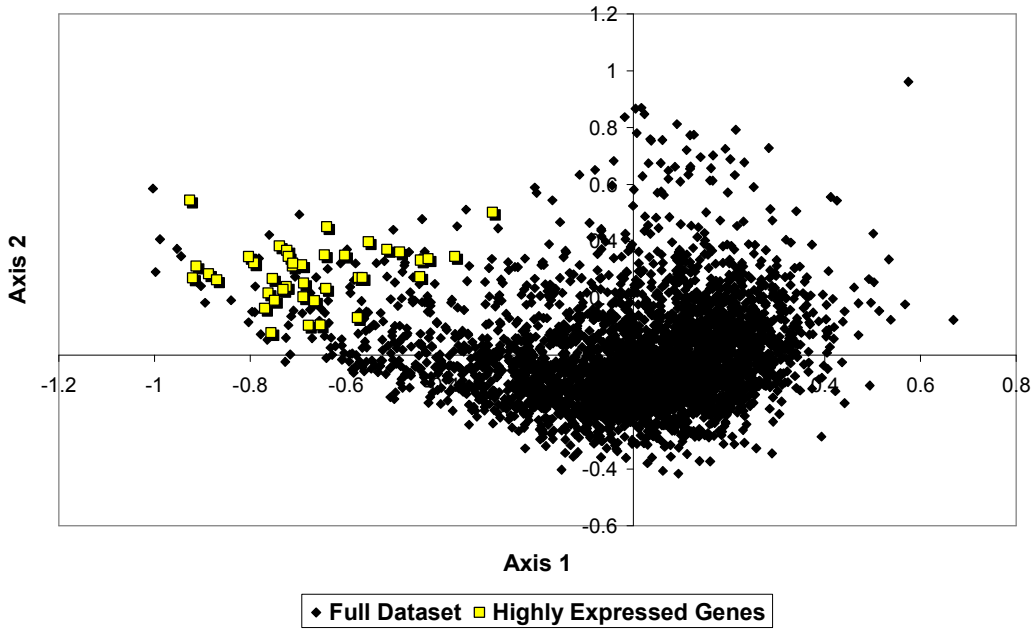

Figure 3-5 Plot of correspondence analysis on RSCU data.
The 40 highly expressed genes (*rplA-F*, *rplI-T*, *rpsB-T*, EF-*Ts*, EF-*Tu* and EF-*G*) are marked in yellow.

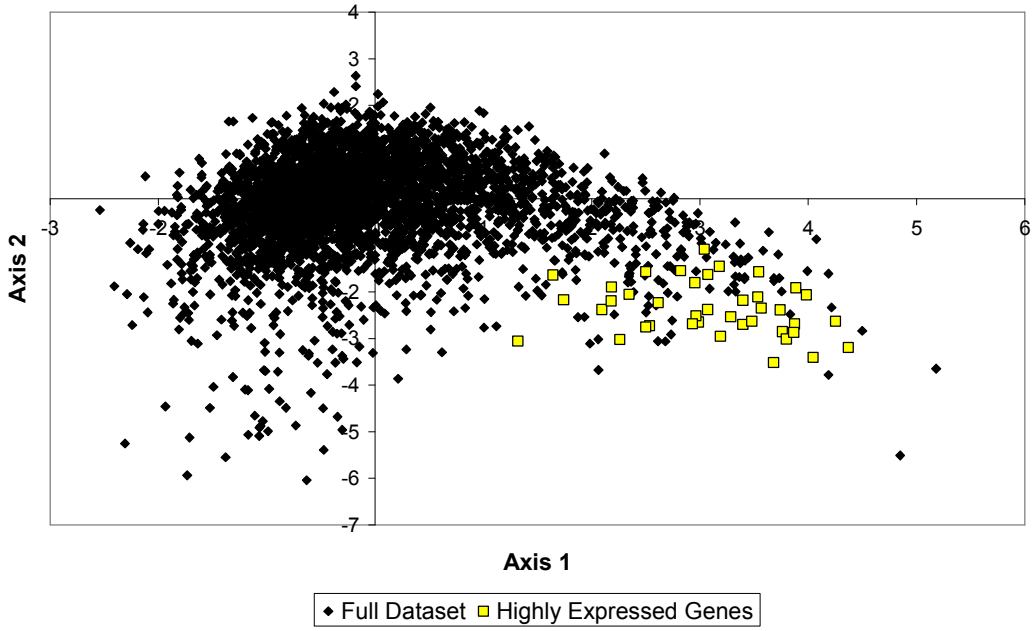

Figure 3-6 Plot of within-block correspondence analysis on raw codon usage data.
The 40 highly expressed genes (*rplA-F, rplI-T, rpsB-T*, EF-*Ts*, EF-*Tu* and EF-*G*) are marked in yellow.

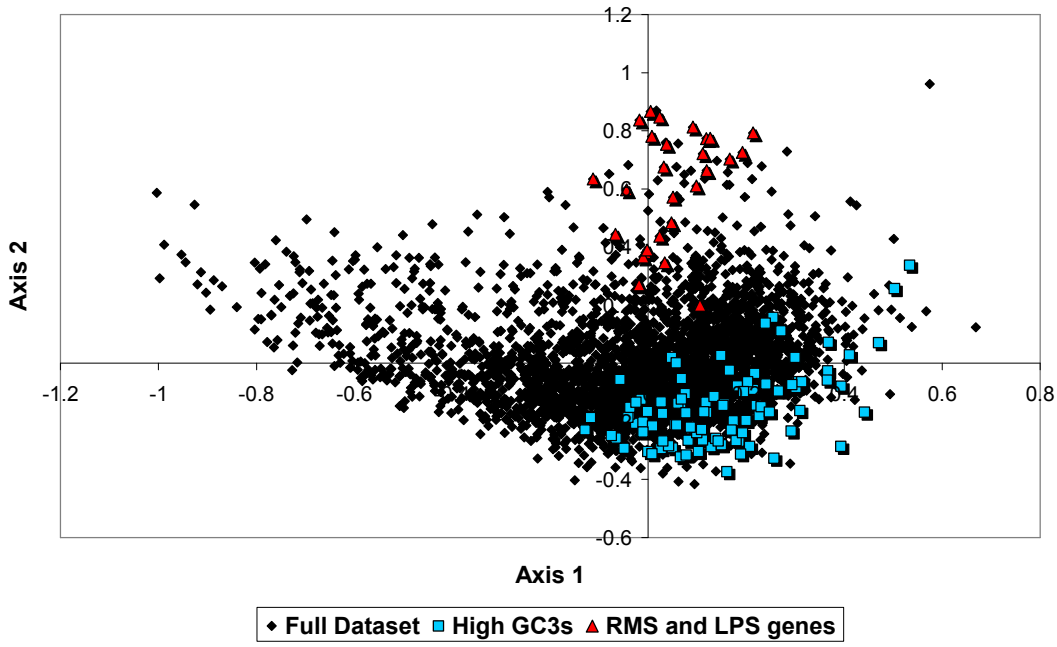

Figure 3-7 Plot of correspondence analysis on RSCU data.
RMS and LPS genes are marked in red whilst genes with the 100 highest GC3s values are marked in blue.

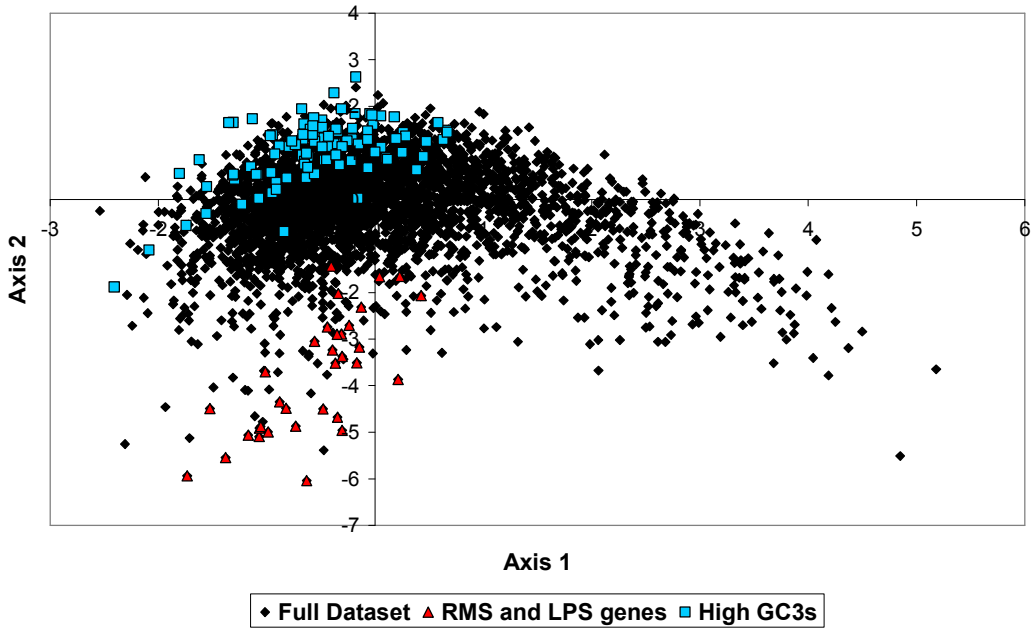

Figure 3-8 Plot of within-block correspondence analysis on raw codon usage data.
RMS and LPS genes are marked in red whilst genes with the 100 highest GC3s values are marked in blue.

rich. These sections were shown to contain lipopolysaccharide (LPS) genes along with restriction modification system (RMS) genes, another largely unidentified collection of genes and a ribosomal gene cluster (section 3.3.1 and Figure 3-2). The ribosomal genes appear to be offset with regard to axis one and axis two, this is probably due to the large number of U and A-ending codons that are potentially optimal codons involved in axis one. The main cause of the A+T rich regions is likely to be horizontally transferred genes. When these genes were *blast*ed (Altschul *et al.*, 1990) they appeared more similar to β and γ proteobacteria orthologues than to anything more closely related and so there is a strong possibility that recent horizontal transfer has taken place (Table 3-3). The cluster of A+T rich lipopolysaccharide genes contains around 21 genes, whilst the restriction modification gene cluster contains between three and six RMS genes. The ribosomal genes were also found to be quite A+T rich, as is often the case, but they are separated from the horizontally transferred genes by axis one. It is extremely unlikely that such crucial genes to an organisms functioning could be horizontally transferred and so these genes are not thought to have arisen from a horizontal transfer event.

3.3.3.1.3 Axis 3 correlated with gene location

When the codons responsible for axis three were examined it was found that usage of G+T or A+C ending codons appeared to be the main cause of the axis. The top 20 codons responsible for the axis were all of the form NNG or NNU (11 G-ending and 9 U-ending) whilst at the other end of the axis the bottom 20 codons were all of the form NNA or NNC (10 A-ending and 10 C-ending). A Pearson correlation between GT3s and axis three was also found to be very strong at 0.80. Axis three accounted for 6.5% of the total variation in the dataset (Table 3-1).

It is a common feature of bacterial genomes that genes on the leading strand are more G+T rich than those on the lagging strand (Rocha *et al.*, 1999). This feature was used previously in this chapter to categorize genes as being either on the leading or lagging strand. When the location of genes as either leading or lagging is overlaid onto axis 3 of the CA a clear separation between the two strands could be easily seen. Leading strand

LPS gene cluster

Bdello Gene		Top Blast hits
Bd1678	probable UDP-glucose 4-epimerase"	• *Pseudomonas aeruginosa (γ)* • *Chromobacterium violaceum (β)* • *Campylobacter coli (β)*
Bd1679	putative aminotransferase"	• *Bacteroides* • *Chromobacterium violaceum (β)*
Bd1680	putative UDP-N-acetylglucosamine-2-epimerase	• *Leptospira* • *Vibrio parahaemolyticus (γ)* • *Vibrio vulnificus (γ)*
Bd1681	putative formyltransferase"	• *Sinorhizobium meliloti (α)* • *Salmonella enterica (γ)* • *Bordetella papertussis (β)*
Bd1683	putative N-acetylneuraminic acid synthetase"	• *Chromobacterium violaceum (β)* • *Vibrio vulnificus (γ)* • *Vibrio parahaemolyticus (γ)*
Bd1684	hexapeptide transferase family protein"	• *Pseudomonas aeruginosa (γ)* • *Caulobacter crescentus (α)* • *Geobacter (δ)*
Bd1685	Mannose-1-phosphate guanyltransferase"	• *Magnetospirillum* • *Leptospira* • *Pseudomonas aeruginosa (γ)*
Bd1686	probable acylneuraminate cytidylyltransferase	• *Leptospira* • *Vibrio vulnificus (γ)* • *Chromobacterium violaceum (β)*
Bd1687	hypothetical protein predicted by	• *Aeromonas punctata (γ)* • *Bacillus cereus* • *Chromobacterium violaceum (β)*
Bd1688	putative polysaccharide biosynthesis protein	• *Aeromonas punctata (γ)* • *Bacillus cereus* • *Chromobacterium violaceum (β)*
Bd1689	hypothetical protein predicted by	• *No Hits*
Bd1690	LPS biosynthesis protein WbpG"	• *Pseudomonas aeruginosa (γ)* • *Vibrio vulnificus (γ)* • *Leptospira*
Bd1691	Imidazole glycerol phosphate synthase subunit	• *Pseudomonas aeruginosa (γ)* • *Vibrio vulnificus (γ)* • *Leptospira*
Bd1692	Imidazole glycerol phosphate synthase subunit	• *Pseudomonas aeruginosa (γ)* • *Leptospira* • *Vibrio vulnificus (γ)*
Bd1693	capsular polysaccharide synthesis enzyme Cap8	• *Pseudomonas aeruginosa (γ)* • *Escherichia coli (γ)*
Bd1694	capsular polysaccharide synthesis enzyme Cap5	• *Pseudomonas aeruginosa (γ)* • *Azoarcus (β)*
Bd1695	capsular polysaccharide synthesis enzyme Cap5	• *Azoarcus (β)* • *Escherichia coli (γ)* • *Pseudomonas aeruginosa (γ)*
Bd1696	putative glycosyltransferase"	• *Dechloromonas aromatica (β)* • *Pseudomonas aeruginosa (γ)* • *Bacteroides*
Bd1697	UDP-N-acetyl-D-quinovosamine 4-epimerase"	• *Fusobacterium nucleatum* • *Vibrio cholerae (γ)* • *Francisella tularensis*
Bd1698	capsular polysaccharide synthesis enzyme Cap8	• *Dechloromonas aromatica (β)* • *Geobacter metallireducens (δ)* • *Polaromonas*
Bd1699	putative acetyltransferase"	• *Caulobacter crescentus (α)* • *Neisseria meningitidis (β)* • *Legionella*

RMS Cluster Genes

Bdello Gene	Description	Top Blast hits
Bd3691	Hypothetical protein	• No Good Hits
Bd3693	Hypothetical protein	• No Good Hits
Bd3694	RMS Gene	• *Methylococcus (γ)* • *Bacillus*
Bd3695	RMS Gene	• *Halobacteria* • *Methanocaldococcus*
Bd3696	RMS Gene	• *Bacillus* • *Shewanella (γ)* • *Desulfovibrio (δ)*
Bd3697	Hypothetical protein	• No Good Hits

Other A+T Rich Region

Bdello Gene	Description	Top Blast hits
Bd2672	Hypothetical protein	• No Good Hits
Bd2673	Hypothetical protein	• No Good Hits
Bd2674	Putative membrane protein	• *Ralstonia* (Not v. good hit)
Bd2675	Putative membrane protein with protease subunit.	• *Ralstonia solanacearum (β)* • *Ralstonia metallireducens (β)* • *Ralstonia eutropha (β)*
Bd2676	Hypothetical protein	• No Good Hits
Bd2677	Hypothetical protein	• No Good Hits
Bd2678	Hypothetical protein	• No Good Hits
Bd2679	Hypothetical protein	• No Good Hits
Bd2680	Hypothetical protein	• No Good Hits
Bd2681	Hypothetical protein	• No Good Hits
Bd2682	Hypothetical protein	• No Good Hits

Table 3-3 Table of top BLAST hits for various potentially horizontally transferred genes.

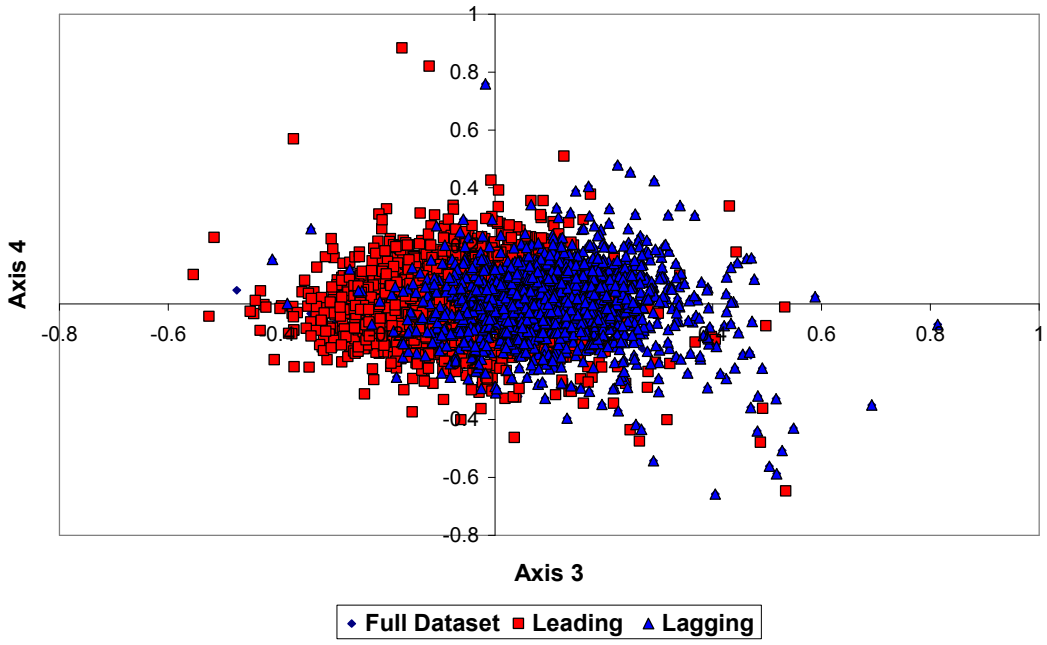

Figure 3-9 Plot of correspondence analysis (axis 3 vs axis 4) on RSCU data.
Leading strand genes are marked in red whilst lagging strand genes are marked in blue.

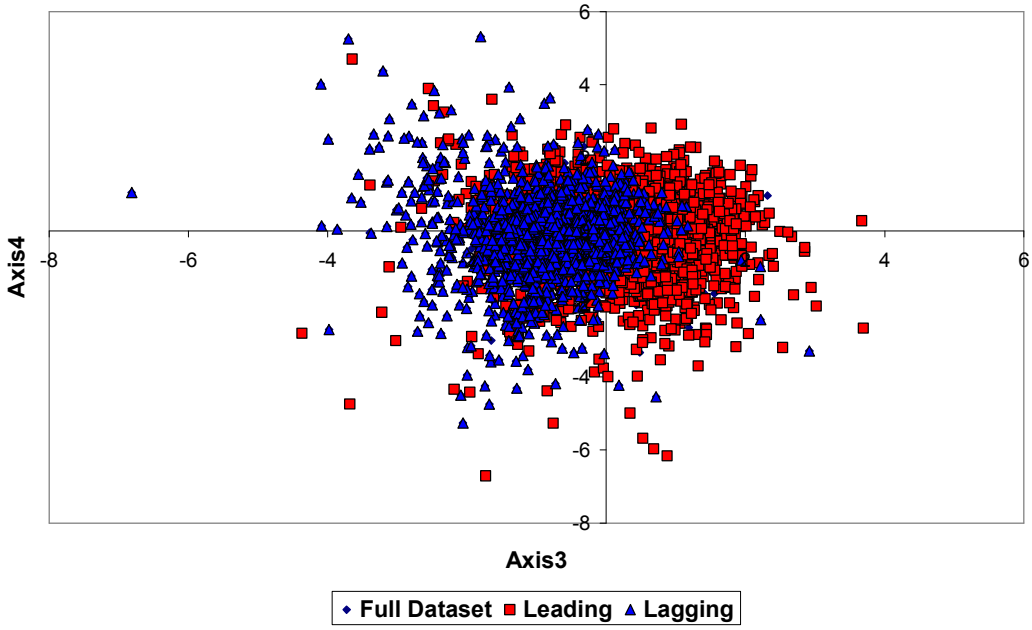

Figure 3-10 Plot of within-block correspondence analysis (axis 3 vs axis 4) on raw codon usage data.
Leading strand genes are marked in red whilst lagging strand genes are marked in blue.

genes were pulled largely to the left hand side and lagging strand genes to the right (Figure 3-9). The mean position for leading strand genes on axis 3 was -0.083 (standard deviation: 0.130), whilst for lagging strand genes it was 0.106 (standard deviation: 0.127).

3.3.3.1.4 Remaining axes

Axis four appears to be correlated with the rare amino acid cysteine. An examination of the codons responsible for axis 4 showed synonymous cysteine codons at either end but this kind of trend has no biological significance. This trend is common when performing CA on RSCU data and is indeed a criticism of using this technique, although as long as one is aware of this problem it can be disregarded in the analysis. Additional axes each contribute very little to the overall variation and are not usually thought to have much biological significance, indeed CodonW only presents data for the first 4 axes. Most correspondence analysis done on other bacterial genomes have only been able to find any biological significance in the trends of the first three or four axes and so the findings here are not surprising.

3.3.3.2 Codon usage analysis using within-block correspondence analysis

The statistical package *R* (http://www.r-project.org/) was used to implement the within-block correspondence analysis using the *ade4* package (Thioulouse *et al.*, 1997) as described in the materials and methods section.

The first three axes produced by both methods identified the same biological features and ranked the influence of such features in the same order. In addition, similarly strong correlations were shown between values such as GC3s and axis two (0.68 for CA on RSCU and 0.78 within-block CA) and GT3s and axis three (0.83 for CA on RSCU and 0.80 for within-block CA). Some of these axes appear inverted with respect to the corresponding axis using the alternative method. However, such a feature has no

consequence in correspondence analysis as the magnitude and direction of the axes are arbitrary and not necessarily comparable to each other.

The remaining axes were again not suitably correlated with anything of biological significance. No correlation between axes four and the use of the rare amino acid cysteine was noticed in contrast to the fourth axis of the correspondence analysis done on the RSCU data. This is an advantage of the within-block correspondence analysis as opposed to the correspondence analysis done on RSCU data as the former weights by amino acid abundance and is thus able to reduce the impact of features due to rare amino acids. However the removal of this axis using within-block CA did not lead to further discovery of factors influencing codon usage bias within this genome.

3.3.4 Genes important in *Bdellovibrio's* unusual predatory lifestyle

3.3.4.1 Defining the optimal codons for *B. bacteriovorus*

Both methods of correspondence analysis, along with the N_C plot, indicated that translational selection was operating in the genome and was indeed the primary cause of codon usage variation.

The 40 ribosomal protein genes seen in the N_C plot and on axis one of both types of CA plot were used as genes known to have high expression levels in other organisms due to their direct involvement in translation. In order to find out which codons were optimal within the genome, codon usage needed to be compared to genes that are believed to be highly expressed and so these 40 genes (*rplA-F*, *rplI-T*, *rpsB-T*, EF-*Ts*, EF-*Tu* and EF-*G*) (Sharp *et al.*, 2005) were used as the standard highly expressed reference set. In order to eliminate the effect of strand bias only leading strand genes were considered in this analysis. The highly expressed gene dataset was now comprised of 37 genes as 3 of the 40 genes were located on the lagging strand. The optimal codons were calculated by comparing leading strand codon usage in the genome as a whole against these 37

AA	Codon	Total	High	AA	Codon	Total	High	AA	Codon	Total	High	AA	Codon	Total	High
Phe	UUU	12964	26	Ser	UCU	8801	140	Tyr	UAU	10831	25	Cys	UGU	3073	14
Phe	UUC	15632	147	Ser	UCC	11832	60	Tyr	UAC	8138	82	Cys	UGC	2878	20
Leu	UUA	2426	6	Ser	UCA	3993	16	TER	UAA	962	26	TER	UGA	363	0
Leu	UUG	17454	166	Ser	UCG	5518	2	TER	UAG	609	9	Trp	UGG	7481	33
Leu	CUU	10798	141	Pro	CCY	5910	78	His	CAU	5294	29	Arg	CGU	11188	216
Leu	CUC	2097	14	Pro	CCC	3952	5	His	CAC	6101	72	Arg	CGC	11566	74
Leu	CUA	1456	19	Pro	CCA	4885	91	Gln	CAA	7808	107	Arg	CGA	1914	0
Leu	CUG	27465	44	Pro	CCG	11357	29	Gln	CAG	17330	51	Arg	CGG	2628	0
Ile	AUU	13476	76	Thr	ACU	8111	197	Asn	AAU	11998	63	Ser	AGU	5975	10
Ile	AUC	19844	239	Thr	ACC	11419	14	Asn	AAC	13336	135	Ser	AGC	7310	40
Ile	AUA	1222	1	Thr	ACA	5604	82	Lys	AAA	24933	442	Arg	AGA	3087	94
Met	AUG	17304	130	Thr	ACG	8196	19	Lys	AAG	17008	172	Arg	AGG	1009	1
Val	GUU	12735	271	Ala	GCU	12505	317	Asp	GAU	18347	107	Gly	GGU	16271	283
Val	GUC	9159	19	Ala	GCC	17536	27	Asp	GAC	14498	118	Gly	GGC	15498	118
Val	GUA	3132	121	Ala	GCA	8080	148	Glu	GAA	26754	208	Gly	GGA	7234	45
Val	GUG	23651	53	Ala	GCG	16541	81	Glu	GAG	14450	110	Gly	GGG	8888	17

Table 3-4 Codon usage for the leading strand of genes on the leading strand of *Bdellovibrio bacteriovorus*.

E. coli Optimal Codons		*B. bacteriovorus* Optimal Codons	
Ala:	GCU, GCA	**Ala:**	GCU, GCA
Arg:	CGU	**Arg:**	CGU, AGA
Asn:	AAC	**Asn:**	AAC
Asp:	GAC	**Asp:**	GAC
Gln:	CAG	**Gln:**	CAA
Glu:	GAA	**Gly:**	GGU
Gly:	GGU	**His:**	CAC
His:	CAC	**Ile:**	AUC
Ile:	AUC	**Leu:**	CUU, UUG
Leu:	CUG	**Lys:**	AAA
Phe:	UUC	**Phe:**	UUC
Pro:	CCG	**Pro:**	CCA, CCU
Ser:	UCU, UCC	**Ser:**	UCU
Thr:	ACU	**Thr:**	ACU, ACA
Tyr:	UAC	**Tyr:**	UAC
Val:	GUU, GUA	**Val:**	GUU, GUA

Table 3-5 Table of optimal codons for *B. bacteriovorus* and *E. coli* genomes.

highly expressed genes (Table 3-4). A chi-squared test was carried out to see which codons were used significantly more in the highly expressed ribosomal genes as compared to the genome as a whole. In addition, a sequential Bonferroni correction (Rice, 1989) was performed to allow for the multiple chi-squared tests. A list of optimal codons was thereby arrived upon (Table 3-5) so that the frequency of optimal codons values for each gene could be calculated.

It is worth noting at this point that these optimal codons match closely with the codons responsible for axis 1 of the correspondence analysis with all of the 21 optimal codons being found amongst the top 23 codons responsible for axis one (Leu:UCC and Ser:CUA were also amongst these codons but not found to be significantly optimal).

3.3.4.2 Difference in putative expression levels of orthologous genes in *B. bacteriovorus* and *E. coli*

In order to look for genes that exhibit highly adapted codon usage the 'frequency of optimal codons' statistic was used (see Chapter 2: Materials and Methods). The frequency of optimal codons is a ratio of optimal codons used in a gene with respect to the total number of synonymous codons in the gene. A gene using only the defined 'optimal' codons would score a value of one whilst one using no such codons would receive a value of zero. It has been seen that genes with highly adapted codon usage are usually highly expressed and as such genes with high F_{OP} values were assumed to be putatively highly expressed. Using the F_{OP} value many potentially highly expressed genes were found in the *B. bacteriovorus* genome. However, it is likely that many of these genes are highly expressed in all genomes with selected codon usage bias. In order to find genes with unique expression patterns, and possibly importance, in *B. bacteriovorus* a comparison organism was needed, preferably a well studied organism with thorough gene annotation and a well understood and more conventional lifestyle.

It was for these reasons that *E. coli* was the obvious choice as the comparison organism. The codon usage of *E. coli* has also been shown

previously to exhibit selected codon usage bias (Ikemura, 1981b). By comparing *Bdellovibrio* genes against their *E. coli* homologues an idea of genes that have a unique importance in the life cycle of *Bdellovibrio* could be established. This meant that the optimal codons in *E. coli* had to be calculated in order for F_{OP} values to be calculated for the *E.coli* genes. This was done by a similar method as for *B. bacteriovorus*, using the *E.* coli orthologues of the same 40 highly expressed genes and the genome as a whole.

In order to compare F_{OP} values, and hence putative expression levels, it was necessary to assign gene orthologue pairs between *B. bacteriovorus* and *E. coli*. Orthologous genes were found using inpararanoid and the HOGENOM database (see section 3.2.1); this approach identified 1061 genes that were then used in the comparison analysis.

To look for genes with uniquely high or low expression levels the F_{OP} values for each of the homologous genes found between *E. coli* and *B. bacteriovorus* were plotted (Figure 3-11). Genes classed as significant were those more than 2 standard errors above or below the model II regression line fitted through the data as described in section 3.2.2.

3.3.4.3 Difference in expression patterns between *B. bacteriovorus* and *E.coli* genes.

The analysis resulted in 133 potential candidate genes for uniquely high levels of expression in *B.bacteriovorus* as compared to *E. coli* and 158 candidate genes for uniquely low levels of expression. These genes are listed in full in appendix A with a summary of the main gene classifications in this chapter (Table 3-6). As can be seen from the table the main categories of genes were remarkably similar but it was not until a closer examination of the major classes of gene in these categories was carried out that a picture began to emerge.

One of the major gene classifications identified was concerned with energy metabolism. It seems that the main glycolysis/gluconeogenesis pathway genes are more highly expressed in *E.coli* than *B. bacteriovorus*. However,

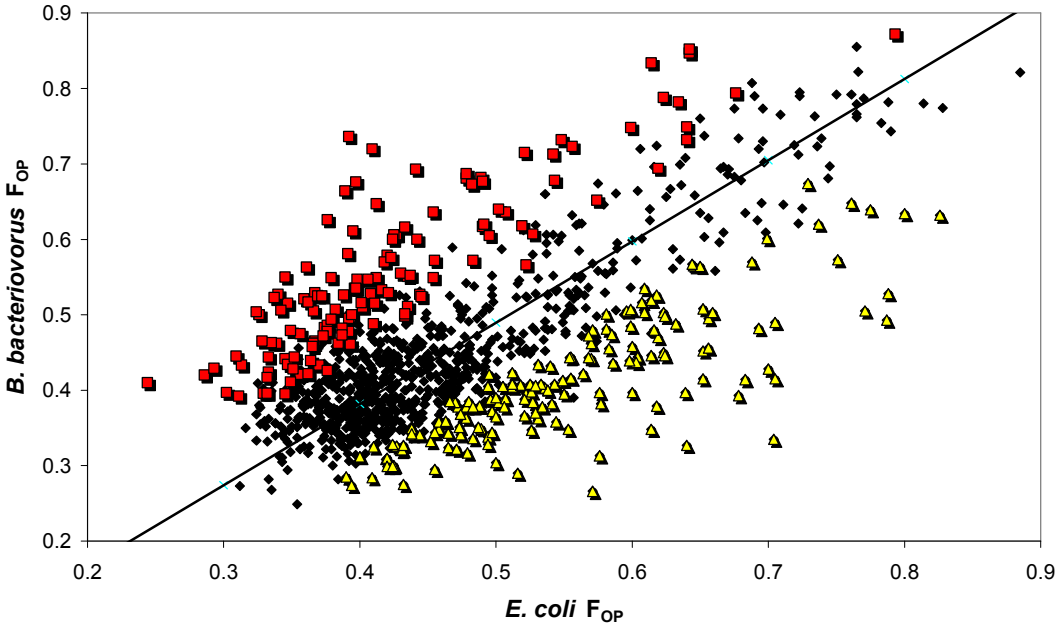

Figure 3-11 A plot of *B. bacteriovorus* vs *E. coli* F_{OP} values.
Points marked in red indicate genes that have significantly high F_{OP} values, more than standard errors (2SE) above the model II regression line. Points marked in yellow indicate genes that have significantly low F_{OP} values, more than standard errors (2SE) below the model II regression line.

Genes with uniquely high F_{OP} in B. bacteriovorus		Genes with uniquely low F_{OP} in B. bacteriovorus	
Energy metabolism	31	Transport and binding proteins	21
Transport and binding proteins	12	Energy metabolism	20
Cellular processes	11	Protein synthesis	20
Cell envelope	9	Purines, pyrimidines, nucleosides, and nucleotides	18
Protein fate	9	Protein fate	13
DNA metabolism	8	Cell envelope	11
Protein synthesis	8	Central intermediary metabolism	6
Fatty acid and phospholipid metabolism	7	DNA metabolism	6
Purines, pyrimidines, nucleosides, and nucleotides	6	Amino acid biosynthesis	5
Regulatory functions	6	Transcription	5
Biosynthesis of cofactors, prosthetic groups, and carriers	4	Cellular processes	4
Amino acid biosynthesis	2	Fatty acid and phospholipid metabolism	4
Central intermediary metabolism	2	Regulatory functions	4
Transcription	2	Biosynthesis of cofactors, prosthetic groups, and carriers	3
Mobile and extrachromosomal element functions	1	Unclassified/Unknown Function	18
Unclassified/Unknown function	15		
Total	**133**	**Total**	**158**

Table 3-6 Table of genes with unique F_{OP} in B. bacteriovorus when compared with *E. coli*

B. bacteriovorus places a stronger emphasis on the TCA cycle and electron transport as well as ATP-synthase genes. Perhaps the increases in these later metabolic pathways are preferred as they produce more ATP than glycolysis, or perhaps pyruvate stores can be salvaged from the host *E. coli* cell. Another major gene classification was concerned with transport and binding proteins. Genes expressed at higher comparative levels in *B. bacteriovorus* were largely concerned with peptide transport whilst those expressed at lower levels were more concerned with small cation and anion transport (such as Na^+, K^+ and Cl^-) as well as phosphates and nitrates. Peptide transport, possibly to transport degradative enzymes into the prey to breakdown host components, appeared to be particularly important to the predator. When genes involved in 'cellular processes' were examined some genes involved in flagella biosynthesis were expressed at higher levels in *Bdellovibrio*, with a few cell division and detoxification proteins significantly under expressed in *Bdellovibrio* in comparison to *E. coli*. The category of protein fate was next investigated. Genes expressed at higher levels in *B. bacteriovorus* included translocases and peptidases whilst genes involved in protein folding and the heat shock response appeared to be less highly expressed. Translocases and peptidases may be important for breaking down the host components at the cost of proteins ensuring correct folding which appear to be much more important to *E. coli*. Next protein synthesis genes were looked at and the main difference in expression noticeable here was that tRNA synthesis genes had higher F_{OP} values in *E. coli* than *B. bacteriovorus*. When genes involved in nucleoside and nucleotide production were examined it appeared that in *Bdellovibrio* genes involved in interconversions were more important than synthesis from scratch which was more important in *E. coli*.

3.4 Discussion

3.4.1 Codon Usage in *Bdellovibrio bacteriovorus*

The primary factor influencing codon usage bias was shown to be selected codon usage bias. This feature was used to determine optimal codons for *B. bacteriovorus* and predict putative expression levels of genes within the

Amino Acid	Codon	Anticodon	tRNA count
Met	AUG	CAT	3
Trp	UGG	CCA	1
Phe	UUU	AAA	
Phe	UUC	GAA	1
Tyr	UAU	ATA	
Tyr	UAC	GTA	1
His	CAU	ATG	
His	CAC	GTG	1
Gln	CAA	TTG	1
Gln	CAG	CTG	
Asn	AAU	ATT	
Asn	AAC	GTT	1
Asp	GAU	ATC	
Asp	GAC	GTC	1
Cys	UGU	ACA	
Cys	UGC	GCA	1
Lys	AAA	TTT	1
Lys	AAG	CTT	
Glu	GAA	TTC	1
Glu	GAG	CTC	
Ile	AUU	AAT	
Ile	AUC	GAT	2
Ile	AUA	TAT	
Pro	CCU	AGG	
Pro	CCC	GGG	1
Pro	CCA	TGG	1
Pro	CCG	CGG	
Thr	ACU	AGT	
Thr	ACC	GGT	1
Thr	ACA	TGT	1
Thr	ACG	CGT	
Val	GUU	AAC	
Val	GUC	GAC	1
Val	GUA	TAC	1
Val	GUG	CAC	
Ala	GCU	AGC	
Ala	GCC	GGC	1
Ala	GCA	TGC	1
Ala	GCG	CGC	
Gly	GGU	ACC	
Gly	GGC	GCC	1
Gly	GGA	TCC	1
Gly	GGG	CCC	

Amino Acid	Codon	Anticodon	tRNA count
Ser	UCU	AGA	
Ser	UCC	GGA	1
Ser	UCA	TGA	1
Ser	UCG	CGA	
Ser	AGU	ACT	
Ser	AGC	GCT	1
Leu	CUU	AAG	
Leu	CUC	GAG	1
Leu	CUA	TAG	1
Leu	CUG	CAG	1
Leu	UUA	TAA	1
Leu	UUG	CAA	1
Arg	CGU	ACG	1
Arg	CGC	GCG	
Arg	CGA	TCG	1
Arg	CGG	CCG	
Arg	AGA	TCT	1
Arg	AGG	CCT	
		Total	36

Table 3-7 Table comparing tRNA abundances with optimal codons, marked in red.

genome. This prediction is based on the hypothesis of translational selection whereby optimal codons are selected for, to ensure efficient and accurate translation of highly expressed genes. This is ensured by using codons that correspond exactly to the most abundant tRNAs in the cell. If translational selection were indeed operating as predicted so far, the optimal codons identified in this chapter should correspond to the tRNA species abundances. To explore this further the genomic tRNA database (http://lowelab.ucsc.edu/GtRNAdb/) was used to explore tRNA abundances in *B. bacteriovorus* and compare them to the optimal codons identified (Table 3-7).

When amino acids with only two possible synonymous codons were considered the correlation between optimal codons and tRNA abundance was particularly strong. The amino acids Phe, Tyr, His, Gln, Asn and Asp all showed a correlation between the tRNA species present and optimal codons, thus agreeing with Ikemura's hypothesis (Ikemura, 1981a; Ikemura, 1981b). The remaining two-fold degenerate amino acids of Cys, Lys and Glu showed no significant optimal codon preference whilst the effectively two-fold degenerate amino acid Ile did show a correlation between tRNA abundance and optimal codon usage. The amino acids with four possible synonymous codons (quartets) all appear to have tRNA species with T and G at their first anticodon positions, thus decoding NNA and NNC codons by exact Crick-Watson base pairing. The amino acids Pro, Thr, Val and Ala all have optimal codons of the form NNA but none have NNC as an optimal codon and instead have NNU. In addition Gly has only one optimal codon, GGU. So accuracy predictions are partially supported for these four-fold degenerate amino acids, with the A-ending optimal codons almost always being correlated with tRNA species using anticodons with T at the first position. However, U-ending codons are preferred even though no exactly complementary tRNA is present and tRNAs with G at the first anticodon position are present with no optimal codon of the form NNC. The amino acids with six possible synonymous codons show a much weaker correlation between tRNA abundances. The NNU codon from the main block of four synonymous codons is optimal for Ser, Arg and Leu although this only corresponds to a tRNA species in Arginine. Whether these patterns of

optimal codons and tRNA abundances are specific to *B. bacteriovorus* or a general feature will be examined further in chapters 5 and 6.

Many optimal codon choices appear to be conserved between *Bdellovibrio* and *E. coli*. In fact, of the 19 *E. coli* optimal codons 13 are also optimal in *B. bacteriovorus* (Table 3-5). For two-fold degenerate amino acids with codons of the form NNY, the NNC codon was optimal in all cases for *Bdellovibrio* and *E. coli* except for cysteine where neither genome showed an optimal codon preference. When amino acids with codons of the form NNR were considered in *E. coli* the amino acid glutamine was preferred for the CAG codon whilst *B. bacteriovorus* preferred the CAA alternative. An optimal codon for lysine was not defined in *E. coli* whilst an optimal codon for glutamate was not defined in *B. bacteriovorus*. The four-degenerate amino acids threonine, valine, alanine and glycine showed similar optimal codon preferences in both genomes with largely NNU or NNU and NNA codon preference (although there was an extra ACA codon preferred for threonine in *B. bacteriovorus* not seen in *E. coli*). Whilst the two species differed for proline with CCG preferred in *E. coli* but CCA and CCU preferred in *Bdellovibrio*. Some similarity was also seen for serine and arginine with UCU and CGU codons preferred in both genomes but additional optimal codons present in *B. bacteriovorus*. Therefore, it seems that some optimal codons may be conserved and common to many species whilst others often differ, this will be investigated further in chapter 5.

Previous work in the Sharp lab has looked at estimating the strength of selected codon usage bias in bacterial genomes (Sharp *et al.*, 2005) and the methods behind this are additionally discussed in chapter two of this thesis. Here it is useful to touch on this work and see how *B. bacteriovorus* fits into the bigger picture. Calculating the strength of selected codon usage bias in *B. bacteriovorus* by comparing codon usage in highly expressed genes, using the 40 highly expressed gene dataset (*rplA-F*, *rplI-T*, *rpsB-T*, EF-*Ts*, EF-*Tu* and EF-*G*) (Sharp *et al.*, 2005), to the genome as a wholes gives an S-value of 1.060 which is lower than *E. coli*'s value of 1.489. If just the leading strand genes are considered in the estimation of the strength of selected codon usage bias, as was done to calculate optimal

codons in *B. bacteriovorus* to minimize the effects of strand bias, an S-value of 1.073 is calculated; a value that is not dramatically different from the initial 1.060 value. However with just 2 rRNA genes and 36 tRNA genes a strength of selected codon usage bias of around 0.5 would be expected rather than one around the 1.0 mark (Sharp *et al.*, 2005). Although *Caulobacter crescentus* has just 2 rRNA genes and an S-value of 1.152 so the value is not totally unreasonable.

3.4.2 Comparing the success of various multivariate analysis methods

The two methods both picked out the same three major axes in the same order and so gave the same conclusions. Axis four in the usual correspondence analysis method was associated with a rare amino acid (Cys) as is often the case with correspondence analysis on RSCU values. The within-block correspondence analysis did not seem to suffer from this flaw but axis four still did not seem to reveal anything useful. Further analysis of the results looking at the correlations of the gene positions along the various axes showed strong correlations between the two methods, showing that they are essentially picking out the same trends. Axes one to three showed very good correlations of 90% or above (Table 3-8) whereas axis four, as expected, showed less than 50% correlation between the two methods.

The within-block method performed using the *ade4* package (Thioulouse *et al.*, 1997) in *R* (http://www.r-project.org/) is essentially a two stage process with the within analysis coming first followed by the correspondence analysis. In order to check the correspondence analysis method implemented in *ade4* was the same as that used in *CodonW* (Peden, 1999) the two methods were also compared. Both methods gave exactly the same results and correlations by axis coordinates and gene rank for the two implementations were 100%. This meant it could be sure that the only difference between the within-block correspondence analysis method and the usual correspondence method was the within analysis.

	Coordinates	Ranks
Axis 1	-0.970	-0.957
Axis 2	0.914	0.902
Axis 3	0.889	0.885
Axis 4	-0.441	-0.437

Table 3-8 Correlations between within-block CA and CA on RSCU data
Correlations are shown both by axis coordinates and by gene ranking order.

It therefore seems that although correspondence analysis on RSCU data does have its draw backs, as long as one is aware what those problems are they are not too serious. As with all forms of multivariate analyses it is the responsibility of the user to assign biological significance to each axis, the method simply looks for trends without assigning cause. However, the within-block correspondence analysis did perform well and in other cases where axes caused by rare codons are more predominant may help to remove this source of error from the analysis.

3.4.3 Differences in expression level for housekeeping genes between *B. bacteriovorus* and *E. coli*

This method of identifying potentially up-regulated genes found many "housekeeping genes" whose expression patterns looked to have been putatively altered in order to adapt to the predatory lifestyle of *Bdellovibrio*. The predation process involves extreme hydrolysis, uptake and resynthesis of macromolecules from prey. This may bring with it extra calls upon the energetic and secretory core machinery of the *Bdellovibrio* cell.

As *Bdellovibrio* replication is at its fastest during predatory, and not host independent, growth (four *Bdellovibrios* can be simultaneously liberated in one hour from a small *E. coli* prey cell, a doubling time of 30 minutes compared to a reported minimum biomass doubling time of 3 hours for host independent growth (Barel & Jurkevitch, 2001)), it is likely that codon optimization relates better to genes used in the predatory growth phase. Such genes are difficult to identify by predatory mutant hunts as they are essential for housekeeping viability roles in *Bdellovibrio* but the method here allowed such housekeeping genes to be identified.

The initial comparison between genes putatively expressed at high levels in *B. bacteriovorus* but not in *E. coli*, and vice versa, gave an insight into some of the changes that a predatory lifestyle has inflicted on *Bdellovibrio*'s housekeeping genes (Table 3-6 and section 3.3.4.3). The emphasis in *Bdellovibrio* seemed to be with peptide transport across membranes, as opposed to cation and anion transport which were more important in *E. coli*.

Also the conversion of nucleosides and nucleotides was seen to be more important in *Bdellovibrio* than *E. coli*, where synthesis from scratch was more emphasized. Protein production in *Bdellovibrio* was more tuned to the rapid production of peptidases and translocases at the cost of chaperones and heat shock response proteins which were seen to be more important in *E.coli*. These changes are indicative of *Bdellovibrio's* unique lifestyle; invading a host cell and remaining in the periplasm whilst transporting degradive enzymes into the host's cytoplasm and scavenging molecules from the host cell rather than synthesizing their own from scratch.

Those genes identified here are currently being tested further experimentally using micro-array analysis and validation studies by the Sockett lab.

3.4.4 Using statistics to estimate gene expression levels

Measures of gene expression levels such as F_{OP} are related to codon usage bias. This means that for genomes with low codon usage bias or a high mutational bias such a method is not informative in predicting gene expression level. Additionally genes that are highly expressed in the cell but not at times critical for optimal growth may not show up using these codon usage bias based prediction methods, due to the nature of the selective pressure on codon usage bias; an example of this is the *metE* gene in *E. coli* (Henry & Sharp, 2007).

The former problem can be solved by performing a simple correspondence analysis, as done here for *Bdellovibrio,* to ascertain that translation selection is indeed the main source of codon usage variation in the organism. If such methods are not done to confirm the main factors affecting intragenomic codon usage patterns then the prediction of highly expressed genes using methods based around synonymous codon usage is entirely meaningless.

Another important aspect of this chapter was the attempt to assess gene expression level from codon usage bias. In a substantial number of papers,

Karlin and colleagues have developed an alternative approach to this. In collaboration with my supervisor I have written a critique of this approach, and described the F_{OP} method used here as a better approach (Henry & Sharp, 2007) (See appendix B).

Chapter 4:
Intergenomic codon usage variation: strength of selected codon usage bias.

4.1 Introduction

4.1.1 Intergenomic vs Intragenomic codon usage patterns

The previous chapter examined variation in codon usage bias in a single genome, that of *Bdellovibrio bacteriovorus*. Similar analyses have been performed on many other bacterial genomes and these studies have elucidated three main factors that are important in affecting intragenomic codon usage in bacteria. These factors are translational selection with genes expressed at high levels using a subset of 'optimal' codons (Ikemura, 1981a; Ikemura 1981b), strand bias favouring a comparatively G+T rich leading strand of replication (Lobry, 1996; McLean *et al.*, 1998) and horizontal transfer resulting in genomic islands of atypical base composition (Ochman *et al.*, 2000). However, the extent to which these factors influence codon usage may vary greatly between genomes. The use of correspondence analysis (Greenacre, 1984) to examine codon usage is useful for an in depth analysis of one particular genome but much work has been done in this area whilst relatively little work has looked at intergenomic codon usage variation. It is for this reason that the remaining chapters of this thesis will now be concerned with the variation in codon usage patterns between bacterial genomes.

There are three main factors influencing intergenomic codon usage patterns among species. Firstly, it has long been evident that the G+C content of bacterial genomes vary drastically (Muto & Osawa, 1987) from extremely G+C rich genomes such as *Micrococcus luteus* (G+C content: 72%) (Ohama *et al.*, 1990) to genomes with a low G+C content such as *Mycoplasma capricolum* (G+C content: 25%) (Ohkubo *et al.*, 1987), with variation in

G+C content at silent sites (GC3s) being even broader. This variation in G+C content is as a result of the differences in mutational biases between bacterial genomes (Chen *et al.*, 2004). Secondly the influence of the GC skew (Lobry, 1996) varies between species with some genomes, such as *Borellia burgdorferi* (Sharp *et al.*, 2005), having genes on their leading strand that are very G+T rich at synonymously variable positions in comparison to lagging strand genes, whilst in other species this feature is less noticeable. Thirdly the strength of selection can vary considerably between species with species such as *Escherichia coli* exhibiting selected codon usage bias (Ikemura, 1981a; Ikemura 1981b) whilst in other species, such as *Helicobacter pylori*, selection does not appear to be a significant influence on codon usage (Lafay *et al.*, 2000).

4.1.2 Variation in strength of selected codon usage bias between bacterial genomes

Many bacterial genomes exhibit a strong degree of translational selection. In these genomes highly expressed genes have a bias towards a subset of synonymous codons, which are those most accurately and/or efficiently recognized by the most abundant tRNA species. The strength of this bias has been shown to be correlated with the level of gene expression (Ikemura, 1985). In the previous chapter translational selection was shown to be the primary factor in the shaping of codon usage bias in the Delta Proteobacterium *B. bacteriovorus*. Similarly, two of the first genomes to be subjected to similar analyses, *Escherichia coli* (Ikemura, 1981a; Ikemura, 1981b; Post & Nomura, 1980) and *Saccharomyces cerevisiae* (Bennetzen & Hall, 1982; Ikemura, 1982), exhibited a high degree of translational selection and many have mistakenly assumed that such selection is present in all unicellular organisms. It is important to realize that in many bacterial genomes translational selection is not a major factor influencing codon usage bias. Indeed codon usage bias in these genomes, such as *Helicobacter pylori* (Lafay *et al.*, 2000), can be much lower so that only the effects of neutral mutation can be seen (Bulmer, 1991).

Relatively little work has been done to examine how and why such variation occurs with only a few previous studies attempting to address this question (dos Reis *et al.*, 2004; Rocha, 2004; Sharp *et al.*, 2005). The paper from dos Reis and colleagues attempted to measure the strength of selected codon usage bias using the effective number of codons in a gene (Wright, 1990) and a modification of the codon adaptation index, termed the tRNA adaptation index. However, Sharp and colleagues found some discrepancies between these estimations and previously published codon usage analyses that were not found using the Sharp method, described later in this section (Sharp *et al.*, 2005). Another attempt by Rocha looked and tRNA species abundances across 102 bacterial species and found that as minimal generation times got shorter, the genomes contained more tRNA genes, but fewer anticodon species; indicating an optimization of the translation machinery to use a small subset of optimal codons and anticodons in fast-growing bacteria and in highly expressed genes.

The work presented here seeks to further the findings from these pieces of research and is, in many respects, a continuation of the work produced by Sharp *et al*., 2005 using the same methods. In this study a method to quantify the strength of selected codon usage bias was devised. This method aimed to overcome two of the major hurdles in comparing codon usage between bacterial species. The first problem was that a suitable method needed to take into account the mutational biases present in the genomes and allow for them. The second was to allow for the fact that the codons considered optimal often vary among species. The method overcame the first problem by using a population genetics model (Bulmer, 1991) and modifying it to take into account background mutational biases. The second problem was overcome by only considering four amino acids (Phe, Tyr, Ile, Asn) where the choice of codon was always between WWU and WWC. Across all bacterial species the WWC codon is preferred as only one tRNA species is present in the genome to decode both codons. The first anticodon position of this tRNA is a guanine and so it pairs exactly with the WWC codon whilst pairing with WWU through wobble, assuming no base modifications occur. This means that WWC is always better

recognized and hence is translationally optimal as it promotes more accurate and efficient translation.

This initial study looked at 80 bacterial genomes and aimed to measure the extent of selected codon usage bias in these species. It was found that the strength of selection was strongly correlated with the number of rRNA operons and tRNA genes present in the genome. It was suggested that the strength of selection was influenced by the generation time of the bacterium (rRNA operon number was used as a surrogate measure of the speed of replication) and so this evidence appeared to indicate that this was the case.

The work presented in this chapter aims to build on this initial study and extend it to incorporate the many newly sequenced complete bacterial genomes that are now available since the original work was done and have caused the dataset to double from the original 80 genomes to 160 genomes in this updated dataset. Newly available information on the generation times of many of these bacteria also allowed more direct analysis of the conclusions of the original paper.

4.2 Materials and Methods

4.2.1 Selecting the dataset

The dataset used was created from GenBank release 148 (15th June 2005, updated Aug 15th 2005). All fully sequenced bacterial genomes were extracted from this GenBank release using the ACNUC interface (Gouy *et al.*, 1985). In order to decrease the amount of redundancy in the dataset, with many of the 'popular' species of bacteria (such as *E. coli*) having multiple strains sequenced, similar strains were removed. To remove similar strains sequence similarity cut offs were used. The method for this was similar to the previous study (Sharp *et al.*, 2005) whereby 5 genes (*rplA, rplB, rplC, rpsB* and *rpsC*) were concatenated and the divergence of this concatenated sequence calculated. This was done in groups by their major taxonomic descriptions, that is to say Gamma Proteobacteria,

Actinobacteria, Firmicutes and so on. Sequences that were more than 3.7% different were kept in the dataset, which excluded all multiple bacterial strains classified in the same species except the three *Buchnera aphidicola* strains (17-26% divergence) and the *Prochlorococcus marinus* strains (25-34% divergence); with the most divergent strains excluded being *Helicobacter pylori* strains 26695 and J99 at 3.3% sequence divergence and the most closely related species pair retained being *Xanthomonas axonopodis* and *Xanthomonas campestris* at 3.7% sequence divergence. This reduced the dataset to 160 bacterial genomes of distinctly different species.

The genes taken to be putatively highly expressed and therefore the most likely to be affected by translational selection were 37 ribosomal protein genes and three elongation factors (*rplA-F*, *rplI-T*, *rpsB-T*, EF-*Ts*, EF-*Tu* and EF-*G*). The only exception to this was the *Mycoplasma penetrans* genome where no *rplI* gene was found so *rplU* was used instead. When finalizing the genes to be included in the highly expressed dataset for each genome a *tBlastn* search (Altschul *et al.*, 1990) was carried out to ensure any copies missed in the *GenBank* sequence annotation could be identified. To do this, the protein sequences of the highly expressed genes from a closely related, and already characterized genome, were queried against a *blast* database created from the genome in question's complete nucleotide sequence. When multiple gene copies were identified, the gene exhibiting the most strongly selected codon usage bias was kept. Another problem involved the GenBank misannotation of a gene's start codon. To check for this each of the 40 genes was separately aligned with those of closely related species, using *ClustalW* (Thompson *et al.*, 1994), to check for obvious errors in start codon position. These checks were carried out to ensure, as much as possible, that the codon usage examined in the putatively highly expressed genes was accurate; this was particularly important considering the relatively few genes in our highly expressed dataset.

4.2.2 Assessing the significance of selection

In order to assess whether the S-values observed were significantly greater than zero an identical method to that in the previous analysis was used

(Sharp *et al.*, 2005). For each genome the set of 40 highly expressed contained on average ~1000 codons considered in the analysis (Phe, Tyr, Ile, Asn). Therefore, datasets were constructed for each genome by adding randomly chosen genes until at least 1000 relevant codons were present. The range of S-values including 95% of these samples was then recorded and if the S-value was greater than the 95% upper limit it was deemed to be significant.

4.2.3 Construction of a phylogeny

Two methods were used for the construction of a phylogenetic tree. The first method concerned the use of 16S rRNA. The sequences were obtained from the *Ribosomal Database Project II release 9* (Olsen *et al.*, 1991) (http://rdp.cme.msu.edu/). As many genomes contain multiple 16s sequences, the majority sequence was always selected. The RDP database provides pre-aligned 16S rRNA sequence aligned using their own alignment algorithm; these pre-aligned sequences were used for all the genomes extracted from the RDP. However, a few genomes did not have any rRNA sequences present in the RDP and instead the sequences were extracted from the *complete microbial resource* at TIGR (http://cmr.tigr.org). These sequences were aligned using *ClustalW* (Thompson *et al.*, 1994) with a profile alignment using a closely related genome sequence with 16S rRNA obtained from the *RDP*. Such a method resulted in a full dataset of 160 16S rRNA sequences, each corresponding to one of the 160 genomes. Tree construction was performed by the bayesian method implemented in *MrBayes* version 3.1.2 (Ronquist & Huelsenbeck, 2003) using the 16S rRNA sequences obtained above and the GTR nucleotide substitution model with gamma distributed rates across the sites. Constraints were set to group major clades, such as the Proteobacteria or Actinobacteria, together to ensure correct phylogenetic grouping as otherwise incorrect major groupings occurred as a result from the large amount of phylogenetic noise in the dataset. These constraints could be placed on the tree prior to construction as implemented in *MrBayes*.

An alternate method was also used which was similar to that used to create the phylogeny for the 80 genome dataset (Sharp *et al.*, 2005). This

method used protein sequences from selected ribosomal protein and elongation factor genes *rplA-C, rpsB-C* and Ef-*Tu*. These genes were aligned individually, then concatenated and gaps removed. The tree was then produced using the same version of *MrBayes* (Ronquist & Huelsenbeck, 2003) using the JTT model of protein evolution and gamma distributed rates with constraints set to keep the proteobacteria together.

4.2.4 Calculation of phylogeny-independent correlations

Due to the highly interrelated nature of the data being examined and the fact that one may expect similar species to have similar genome characteristics, such as G+C content or rRNA operon number, as a result of their close evolutionary distance, a method of producing phylogeny independent correlations was needed. Such a method is implemented in a piece of software called *Continuous* (Pagel, 1999). This program uses the generalized least squares approach for the across-species analysis of comparative data to ensure that correlations among species characters are phylogeny independent. The program was supplied with the phylogeny created using the ribosomal protein sequence data (Figure 4-3) to carry out this task.

4.3 Results

4.3.1 Bacterial genomes dataset

The strength of selected codon usage bias was measured for an additional 80 genomes making the final dataset 160 bacterial genomes in total (Table 4-1). The addition of so many new species reflects the huge rate at which genomes are being sequenced. The addition of these 80 new genomes again followed a stringent 3.7% sequence similarity cut-off value to exclude strains that were too closely related (section 4.2.1). The only genomes still present with multiple strains were the multiple *Buchnera aphidicola* and *Prochlorococcus marinus* strains which are more different from each other than many species. The *Buchnera aphidicola* strains were more different from each other than many organisms classed as distinct species with sequence diveregence of 17-26% between the three strains. This was also

true for the *Prochlorococcus marinus* strains with sequence divergence ranging from 25-34% between the four strains. The most divergent strains excluded were *Helicobacter pylori* strains 26695 and J99 at 3.3% sequence divergence. This 3.7% cut-off also led to the exclusion of *Shigella flexneri* (0.2% different from *E. coli* K12), *Yersinia pseudotuberculosis* (<0.05% different from *Yersinia pestis*), *Neisseria gonorrhoeae* (3.3% different from *Neisseria meningitidis*), *Rickettsia felis* (2.6% different from *Rickettsia conorii*), *Brucella suis* (0.3% different from *Brucella melitensis*), *Brucella abortus* (0.2% different from *Brucella melitensis*), *Bacillus thuringiensis* (<0.05% from *Bacillus anthracis*), *Bacillus cereus* (0.1-1.7% different from *Bacillus anthracis* depending on strain chosen), *Listeria innocua* (1.5% different from *Listeria monocytogenes*) and *Mycobacterium bovis* (0.1% different from *Mycobacterium tuberculosis*). The most closely related species pair retained were *Xanthomonas axonopodis* and *Xanthomonas campestris* at 3.7% sequence divergence. These criteria reduced the dataset to 160 bacterial genomes with distinctly different species.

The phylogeny was now even more diverse than it had been previously (Figure 4-3, Figure 4-4) with the addition of new groups such as the Delta Proteobacteria. Additionally, the resolution was increased with more species represented in each major clade allowing a much more comprehensive analysis of selected codon usage bias across bacterial genomes. The new genomes added were widely distributed across the original phylogeny but were particularly useful in adding more resolution to underrepresented clades. The Alpha Proteobacteria gained more resolution to its G+C-poor Rickettsiales clade, whilst the Beta Proteobacteria also increased from just three species to nine species. The Gamma Proteobacteria and Firmicutes were already well represented but were even more so in this new larger dataset. The Delta Proteobacteria had no representation at all in the original dataset but now have four species whilst the epsilon proteobacteria doubled in size from two to four represented genomes. The resolution of the Actinobacteria was also greatly enhanced with the addition of seven new genomes to take the total to 15 genomes. The smaller groups such as the Spirochaetes, Chlamydiales, Bacteroidetes and Cyanobacteria now also had a much greater resolution.

Species code[a]	rRNA[b]	tRNA[c]	ORF[d]	GC3s[e]	S[f]	Lower[g]	Upper[h]	Gen Time[i]	Accession Number[j]	Species
Alpha Proteobacteria										
Agrtum	4	53	4661	71.0	1.048	-0.202	0.217	3.0	AE008688*	Agrobacterium tumefaciens C58 (UW)
Anamar	1	37	949	51.1	-0.083	-0.175	0.168		CP000030	Anaplasma marginale
Barhen	2	44	1612	26.8	-0.373	-0.311	0.276	3.0	BX897699	Bartonella henselae
Barqui	2	44	1308	27.6	-0.315	-0.316	0.290	3.0	BX897700	Bartonella quintana
Brajap	1	50	8317	82.0	0.741	-0.281	0.312	20.0	BA000040	Bradyrhizobium japonicum
Brumel	3	54	3198	66.0	0.896	-0.202	0.237	2.0	AE008917*	Brucella melitensis
Caucre	2	51	3737	86.0	1.152	-0.310	0.370	1.5	AE005673	Caulobacter crescentus
Ehrcan	1	36	925	16.5	-0.765	-0.487	0.397		CP000107	Ehrlichia canis Jake
Ehrrum	1	36	920	15.8	-0.673	-0.434	0.381		CR767821	Ehrlichia ruminatium strain Welgevonden
Gluoxy	4	50	2432	71.7	0.785	-0.316	0.323	1.0	CP000009	Gluconobacter oxydans
Meslot	2	50	6752	79.0	0.757	-0.245	0.283		BA000012	Mesorhizobium loti
Pelubi	1	32	1354	14.6	0.353	-0.248	0.228		CP000084	Pelagibacter ubique
Rhopal	2	49	4833	83.4	0.505	-0.294	0.324	9.0	BX571963	Rhodopseudomonas palustris
Riccon	1	33	1374	21.0	-0.410	-0.214	0.234	4.1	AE006914	Rickettsia conorii
Ricpro	1	33	834	16.0	-0.421	-0.243	0.225	10.0	AJ235269	Rickettsia prowazekii
Rictyp	1	33	838	16.0	-0.460	-0.258	0.215	10.0	AE017197	Rickettsia typhi
Silpom	3	53	3810	80.3	0.925	-0.224	0.273		CP000031	Silicibacter pomeroyi
Sinmel	3	54	6205	79.0	0.637	-0.225	0.236	1.5	AL591688	Sinorhizobium Meliloti
Wolpip	1	34	1195	25.7	-0.684	-0.197	0.174		AE017196	Wolbachia pipientis
Woltrs	1	34	805	25.6	-0.574	-0.186	0.180	14.0	AE017321	Wolbachia strain TRS
Zymmob	3	51	2001	43.0	0.750	-0.238	0.238		AE008692	Zymomonas mobilis ZM4
Beta Proteobacteria										
Azoebn	4	58	4128	81.6	-0.055	-0.336	0.372	4.3	CR555306	Azoarcus sp. EbN1
Borper	3	51	3804	87.9	-0.033	-0.258	0.291	6.0	BX470248	Bordertella pertussis
Burpse	4	61	5855	87.7	0.340	-0.382	0.369	0.7	BX571965*	Burkholderia pseudomallei
Chrvio	8	98	4407	85.3	0.545	-0.588	0.569	0.8	AL646052*	Chromobacterium violaceum
Decaro	4	64	4171	71.4	0.323	-0.313	0.339		CP000089	Dechloromonas aromatica
Neimen	4	58	2121	60.0	-0.099	-0.346	0.373	1.0	AL157959	Neisseria meningitidis Z2491
Niteur	1	41	2574	53.0	-0.884	-0.253	0.258	18.5	AL954747	Nitrosomonas europaea
Raleut	5	65	5846	81.5	0.675	-0.246	0.282		CP000090*	Ralstonia eutropha JMP134
Ralsol	4	57	5120	87.0	0.024	-0.371	0.451	4.0	AL646052*	Ralstonia solanacearum
Gamma Proteobacteria										
Aciadp	7	76	3324	31.4	1.545	-0.266	0.270	0.5	CR543861	Acinetobacter sp.ADP1
Bloflo	1	37	589	13.0	-1.067	-0.625	0.491	36.0	BX248583	Blochmannia floridandus
Blopen	1	39	610	17.2	-0.074	-0.290	0.220		CP000016	Blochmannia pennsylvanicus
Buchap	1	32	564	12.0	-0.017	-0.228	0.179		BA000003	Buchnera aphidicola Ap
Buchbp	1	32	504	12.0	-0.590	-0.448	0.356		AF492592	Buchnera aphidicola Bp
Buchsg	1	32	545	10.0	-0.069	-0.265	0.213	36.0	AE013218	Buchnera aphidicola Sg
Colpsy	9	88	4910	27.8	1.344	-0.214	0.206		CP000083	Colwellia psychrerythraea
Coxbur	1	42	2009	38.0	0.175	-0.184	0.170	8.0	AE016828	Coxiella burnetti
Erwcar	7	76	4492	54.9	0.951	-0.249	0.272	0.2	BX950851	Erwinia carotovora
Esccol	7	86	4289	54.0	1.489	-0.286	0.308	0.3	U00096	Escherichia coli K-12
Fratul	3	38	1804	19.2	0.562	-0.243	0.252		AJ749949	Francisella tularensis
Haeduc	6	45	1717	27.1	0.937	-0.326	0.252	1.8	AE017143	Haemophilus ducreyi
Haeinf	6	54	1709	27.0	1.492	-0.325	0.330	0.5	L42023	Haemophilus influenzae
Idiloi	4	56	2628	45.0	1.152	-0.207	0.236		AE017340	Idiomarina loihensis L2TR
Legpne	3	43	2943	30.6	0.101	-0.213	0.197	3.3	AE017354	Legionella pneumophila Philadelphia 1

Species code[a]	rRNA[b]	tRNA[c]	ORF[d]	GC3s[e]	S[f]	Lower[g]	Upper[h]	Gen Time[i]	Accession Number[j]	Species
Mansuc	6	60	2385	39.9	1.192	-0.283	0.274		AE016827	Mannheimia succiniciproducens
Metcap	2	46	2960	79.6	-0.265	-0.287	0.308		AE017282	Methylococcus capsulatus
Pasmul	6	57	2014	32.0	1.339	-0.282	0.289	1.0	AE004439	Pasturella multocida
Pholum	7	85	4905	38.6	1.034	-0.299	0.332	0.5	BX470251	Photohabdus luminescens
Phopro	15	168	5414	34.4	1.535	-0.248	0.289	2.5	CR354531*	Photobacterium profundum
Pseaer	4	63	5566	87.0	-0.019	-0.507	0.484	0.5	AE004091	Pseudomonas aeruginosa
Pseflu	5	71	6137	81.9	0.452	-0.324	0.364		CP000076	Pseudomonas fluorescens Pf-5
Pseput	7	73	5350	77.0	0.917	-0.317	0.360	1.1	AE015451	Pseudomonas putida
Psesyr	5	63	5566	71.0	0.701	-0.243	0.255	1.5	AE016853	Pseudomonas syringae
Psyarc	3	49	2147	38.6	1.037	-0.236	0.251		CP000082	Psychrobacter arcticum
Salent	7	85	4452	58.0	1.522	-0.254	0.292	0.4	AE006468	Salmonella enterica
Sheone	9	102	4630	45.0	1.377	-0.275	0.313	0.7	AE014299	Shewenella oneidensis
Vibcho	8	98	3828	47.0	1.725	-0.273	0.294	0.2	AE003852*	Vibrio cholerae
Vibfis	12	119	3744	26.1	2.001	-0.347	0.348	0.3	CP000020*	Vibrio fischeri
Vibpar	11	126	4832	44.0	1.886	-0.300	0.336	0.2	BA000031*	Vibrio parahaemolyticus
Vibvul	9	112	4959	47.0	1.950	-0.266	0.296	0.2	AE016795*	Vibrio vulnificus CMCP6
Wigbre	2	34	611	9.0	0.105	-0.247	0.203		BA000021	Wigglesworthia glossinidia
Xanaxo	2	54	4312	80.0	0.636	-0.261	0.273	7.0	AE008923	Xanthomonas axonopodis
Xancam	2	53	4181	81.0	0.607	-0.299	0.292	3.0	AE008922	Xanthomonas campestris
Xanory	2	54	4640	77.1	0.535	-0.219	0.283	2.0	AE013598	Xanthomonas oryzae
Xylfas	2	49	2034	54.0	-0.781	-0.324	0.382	96.0	AE009442	Xylella fastidiosa Temecula
Yerpes	6	70	4008	48.0	1.153	-0.243	0.258	1.2	AL590842	Yersinia pestis CO92
Delta Proteobacteria										
Bdebac	2	36	3583	56.9	1.060	-0.279	0.300	1.4	BX842601	Bdellovibrio bacteriovorus
Despsy	7	64	3118	46.5	0.056	-0.349	0.326	0.4	CR522870	Desulfotalea psychrophila
Desvul	5	68	3379	77.3	0.473	-0.436	0.443	14.0	AE017285	Desulfovibrio vulgaris
Geosul	2	49	3448	77.0	-0.384	-0.327	0.340	6.0	AE017180	Geobacter sulfurreducens
Epsilon Proteobacteria										
Camjej	3	43	1654	17.0	0.486	-0.375	0.300	1.5	AL111168	Campylobacter jejuni 11168
Helhep	1	37	1876	25.3	0.019	-0.309	0.247	4.2	AE017125	Helicobacter hepaticus
Helpyl	2	36	1491	41.0	0.016	-0.195	0.184	2.4	AE001439	Helicobacter pylori J99
Wolsuc	3	40	2047	53.7	0.563	-0.219	0.225	1.0	BX571656	Wolinella succinogenes
Firmicutes										
Bacant	11	95	5311	23.0	2.045	-0.316	0.338		AE016879	Bacillus anthracis Ames
Baccla	7	74	4108	42.0	0.767	-0.175	0.178		AP006627	Bacillus clausii
Bachal	8	78	4066	40.0	0.999	-0.174	0.166	0.6	BA000004	Bacillus halodurans
Baclic	7	72	4152	50.1	1.072	-0.216	0.196	0.6	CP000002	Bacillus licheniformis strain ATCC 14580
Bacsub	10	88	4100	43.0	1.360	-0.224	0.232	0.4	AL009126	Bacillus subtilis
Cloace	11	73	3672	18.0	0.838	-0.286	0.283	0.6	AE001437	Clostridium acetobutylicum
Cloper	10	96	2660	14.0	2.648	-0.420	0.434	0.2	BA000016	Clostridium perfringens
Clotet	6	54	2373	14.0	1.004	-0.272	0.244		AE015927	Clostridium tetani
Entfae	4	68	3133	28.0	1.840	-0.287	0.324	0.5	AE016830	Enterococcus faecalis
Geokau	9	87	3498	60.5	0.559	-0.292	0.247		BA000043	Geobacillus kaustophilus
Lacaci	4	61	1866	22.2	1.361	-0.394	0.347	1.8	CP000033	Lactobacillus acidophilus
Lacjoh	6	79	1822	21.6	1.502	-0.325	0.340	0.9	AE017198	Lactobacillus johnsonii
Laclac	6	62	2266	23.0	2.288	-0.321	0.334	0.7	AE005176	Lactococcus lactis lactis
Lacpla	5	70	3051	43.0	1.253	-0.268	0.271	1.6	AL935263	Lactobacillus plantarum
Lismon	6	67	2855	28.0	1.198	-0.288	0.296	1.0	AL591824	Listeria monocytogenes EGD
Mesflo	2	29	683	10.8	1.418	-0.304	0.346		AE017263	Mesoplasma florum

Species code[a]	rRNA[b]	tRNA[c]	ORF[d]	GC3s[e]	S[f]	Lower[g]	Upper[h]	Gen Time[i]	Accession Number[j]	Species
Mycgal	2	33	726	22.0	0.498	-0.391	0.285	1.0	AE015450	Mycoplasma gallisepticum
Mycgen	1	33	480	22.0	0.318	-0.310	0.269	12.0	L43967	Mycoplasma genitalium
Mychyo	1	30	691	18.4	0.101	-0.221	0.222		AE017332	Mycoplasma hyopneumoniae strain 232
Mycmob	1	28	635	9.7	0.434	-0.295	0.209		AE017308	Mycoplasma mobile
Mycmyc	2	30	1017	7.8	0.650	-0.454	0.372		BX293980	Mycoplasma mycoides
Mycpen	1	30	1037	12.0	0.496	-0.253	0.237		BA000026	Mycoplasma penetrans
Mycpne	1	33	688	41.0	0.324	-0.217	0.206	6.0	U00089	Mycoplasma pneumoniae
Mycpul	1	29	782	13.0	0.380	-0.267	0.235	2.0	AL445566	Mycoplasma pulmonis
Mycsyn	2	34	672	15.2	0.636	-0.393	0.317		AE017245	Mycoplasma synoviae
Oceihe	7	69	3496	23.0	1.301	-0.197	0.180		BA000028	Oceanobacillus iheyensis
Phyast	2	32	755	18.2	0.218	-0.564	0.383		AP006628	Phytoplasma asteris OY
Staaur	5	62	2593	20.0	1.564	-0.267	0.248	0.4	BA000018	Staphylococcus aureus N315
Staepi	5	60	2419	19.0	1.164	-0.243	0.254	0.8	AE015929	Staphylococcus epididermis
Stahae	5	59	2678	19.5	1.442	-0.273	0.259		AP006716	Staphylococcus haemolyticus
Stasap	6	61	2446	20.4	1.355	-0.256	0.247		AP008934	Staphylococcus saprophyticus
Straga	7	80	2124	23.0	1.504	-0.252	0.282	1.8	AE009948	Streptococcus agalactiae 2603V/R
Strpne	4	58	2043	34.0	1.720	-0.364	0.380	0.5	AE007317	Streptococcus pneumoniae R6
Strpyo	6	61	1696	30.0	1.759	-0.286	0.299	0.4	AE004092	Streptococcus pyogenes M1 GAS SF370
Strthe	6	67	1889	31.4	1.656	-0.315	0.363	0.4	CP000023	Streptococcus thermophilus LMG 18311
Symthe	6	98	3337	90.4	-0.150	-0.352	0.421	4.2	AP006840	Symbiobacterium thermophilum
Theten	4	55	2588	32.0	0.457	-0.266	0.265	1.1	AE008691	Thermoanaerobacter tengcongensis
Ureure	2	30	611	11.0	0.401	-0.262	0.232	0.9	AF222894	Ureaplasma urealyticum
Actinobacteria										
Biflon	4	56	1729	75.0	1.343	-0.449	0.519		AE014295	Bifidobacterium longum
Cordip	5	54	2320	54.6	1.861	-0.365	0.384		BX248353	Corynebacterium dipheriae
Coreff	5	56	2950	79.0	1.039	-0.395	0.495		BA000035	Corynebacterium efficiens
Corglu	6	60	3099	58.0	2.185	-0.381	0.467	1.2	BA000036	Corynebacterium glutamicum
Corjej	3	50	2104	75.7	1.588	-0.290	0.376		CR931997	Corynebacterium jejkeium
Leixyl	1	45	2030	86.7	0.522	-0.383	0.459	5.0	AE016822	Leifsonia xyli
Mycavi	1	46	4350	89.3	1.184	-0.297	0.389	10.0	AE016958	Mycobacterium avium
Myclep	1	47	2720	64.0	0.515	-0.193	0.224	240.1	AL450380	Mycobacterium leprae
Myctub	1	45	3918	79.0	0.453	-0.242	0.256	24.0	AL123456	Mycobacterium tuberculosis
Nocfar	3	53	5683	91.0	1.413			3.0	AP006618	Nocardia farcinica
Proacn	3	45	2297	53.1	0.621	-0.250	0.284	5.1	AE017283	Propionibacterium acnes
Strave	6	68	7575	91.0	0.686	-0.501	0.703		BA000030	Streptomyces avermitilis
Strcoe	6	63	7825	93.0	0.987	-0.618	1.049	2.2	AL645882	Streptomyces coelicolor
Thefus	4	52	3110	84.6	0.439	-0.396	0.494		CP000088	Thermobifida fusca
Trowhi	1	49	808	41.0	0.014	-0.191	0.189		AE014184	Tropheryma whipplei Twist
Cyanobacteria										
Glovio	1	45	4430	76.4	0.370	-0.267	0.256		BA000045	Gloeobacter violaceous
Nostoc	4	67	5366	33.0	0.763	-0.271	0.295		BA000019	Nostoc sp. PCC7120
Pro137	1	40	1882	22.4	0.044	-0.253	0.217	17.0	AE017126	Prochlorococcus marinus marinus CCMP1375
Promed	1	37	1716	17.4	0.445	-0.258	0.233		BX548174	Prochlorococcus marinus pastoris CCMP1986 MED4
Promit	2	43	2273	49.6	0.715	-0.309	0.284		BX549175	Prochlorococcus marinus strain MIT9313
Pronat	1	38	1890	21.8	0.433	-0.258	0.276		CP000095	Prochlorococcus marinus strain NATL2A
Sy6803	2	42	3056	48.0	0.616	-0.253	0.243		BA000022	Synechocystis PCC6803
Synelo	2	45	2525	59.1	0.776	-0.279	0.271		AP008231	Synechococcus elongatus PCC6301

Species code[a]	rRNA[b]	tRNA[c]	ORF[d]	GC3s[e]	S[f]	Lower[g]	Upper[h]	Gen Time[i]	Accession Number[j]	Species
Synspp	2	43	2526	68.3	0.918	-0.410	0.412	6.0	BX548020	Synechococcus sp. WH8102
Theelo	1	42	2475	57.0	0.178	-0.207	0.306		BA000039	Thermosynechococcus elongatus
Spirochaetes										
Borbur	1	34	850	19.0	-0.308	-0.579	0.436	4.0	AE000783	Borrelia burgdorferi
Borgar	2	31	832	18.0	-0.206	-0.527	0.412		CP000013	Borrelia garinii
Lepint	1	37	4358	37.0	0.670	-0.258	0.254	9.0	AE010300*	Leptospira interrogans Lai
Treden	2	44	2767	32.7	0.620	-0.333	0.304	5.0	AE017226	Treponema denticola
Trepal	2	45	1031	53.0	-0.015	-0.255	0.248	33.0	AE000520	Treponema pallidum
Chlamydiales										
Chlabo	1	38	961	31.8	-0.148	-0.209	0.243	24.0	CR848038	Chlamydophila abortus
Chlcav	1	38	998	30.0	0.113	-0.208	0.224	24.0	AE015925	Chlamydophila caviae
Chlmur	1	37	904	40.0	0.145	-0.239	0.244		AE002160	Chlamydia muridarum
Chlpne	1	38	1110	33.0	-0.065	-0.234	0.223	24.0	AE002161	Chlamydophila pneumoniae AR39
Chltra	2	37	894	41.0	0.132	-0.247	0.236	24.0	AE001273	Chlamydia trachomatis
Parspp	3	35	2031	24.5	0.347	-0.213	0.218	48.0	BX908798	Parachlamydia sp. UWE25
Bacteroidetes/Chlorobi										
Bacfra	6	74	4578	43.5	0.383	-0.381	0.385	0.6	AP006841	Bacteroides fragilis
Bacthe	5	71	4778	43.0	0.237	-0.418	0.445	1.5	AE015928	Bacteroides thetaiotamicron
Chltep	2	50	2252	72.0	0.069	-0.311	0.301	2.0	AE006470	Chlorobium tepidum
Porgin	4	53	1909	68.3	0.021	-0.305	0.303	2.7	AE015924	Porphyromonas ginigivalis
Fusobacteria										
Fusnuc	5	47	2067	10.0	1.244	-0.274	0.242	0.7	AE009951	Fusobacterium nucleatum
Aquifex										
Aquaeo	2	44	1522	47.0	0.393	-0.273	0.260		AE000657	Aquifex aeolicus
Thermotaogae										
Themar	1	46	1846	51.0	0.365	-0.276	0.281	1.2	AE000512	Thermotoga maritima
Deinococcus-Thermus										
Deirad	3	49	2936	84.0	1.491	-0.280	0.299		AE000513*	Deinococcus radiodurans
Thethe	2	47	1982	92.0	-0.158	-0.422	0.584	2.5	AE017221	Thermus thermophilus
Chloroflexi										
Deheth	1	46	1580	51.4	0.063	-0.200	0.210	19.0	CP000027	Dehalococcoides ethenogenes
Planctomycetes										
Rhobal	1	70	7325	60.3	0.825	-0.222	0.275	10.0	BX119912	Rhodopirellula baltica

Table 4-1 The 160 genome dataset used including relevant genome attributes discussed in this chapter.

[a]The species code as used in Figures 4-2, 4-3 and 4-4

[b]The number of rRNA operons present in the genome obtained from Genbank

[c]The number of tRNA genes present in the genome obtained from the genomic tRNA database (http://lowelab.ucsc.edu/GtRNAdb/)

[d]The number of open reading frames, as obtained from GenBank

[e]The genomic GC3s value

[f]The calculated strength of selected codon usage bias

[g+h]The 95% range of values of 'S' among 1000 sets of randomly selected genes

[i]The GenBank accession number for the genome sequence; asterisk indicates species with two chromosomes; the accession number for the two chromosomes are consecutive, except for *D. radiodurans*, where the second accession number is AE001825.

4.3.2 Strength of selected codon usage bias

One of the problems to be overcome in looking at selected codon usage bias across many genomes was that variation in G+C content of the genomes had to be accounted for. The method used here was devised to do this (Sharp *et al.*, 2005) and appears to have been successful as there is no correlation between 'S' and G+C content despite the wide ranging G+C contents of the 160 genomes in the dataset (Figure 4-1). However, there may be a small problem with very G+C rich genomes, where it is more troublesome assessing the significance of the S-value. In addition, there were a few cases when unusual intragenomic variation in G+C content was found to affect the estimation of 'S' but these cases are discussed later in this chapter. The graph here shows that intergenomic variation in G+C content was largely eliminated so that the effects of neutral mutation and selection could be separated.

The S-values calculated to measure the strength of selected codon usage bias within individual genomes ranged widely from a maximum of 2.65 for *Clostridium perfringens* to a minimum of -1.07 for *Blochmannia floridandus*, with an average 'S' value of 0.64 across the 160 genome dataset. In the dataset as a whole, 105 of the 160 genomes (66%) showed significant evidence of positive selected codon usage bias with 50 of the 160 genomes (31%) shown to be strongly influenced by selection with 'S' values of 1.00 or more. This proportion remained relatively constant during the increase in size of the dataset with the initial dataset showing 24 from 80 genomes with no significant selection (30%) and the increased dataset showing 42 from 160 genomes with no evidence of selection (26%). In addition 13 bacterial species (8%) showed evidence of significant negative selection. The issue of significant negative S-values is addressed in the discussion section of this chapter.

The addition of the new genomes failed to increase the proportion of genomes displaying significantly positive selection in clades such as the Beta Proteobacteria where just a single species has a significant positive S-

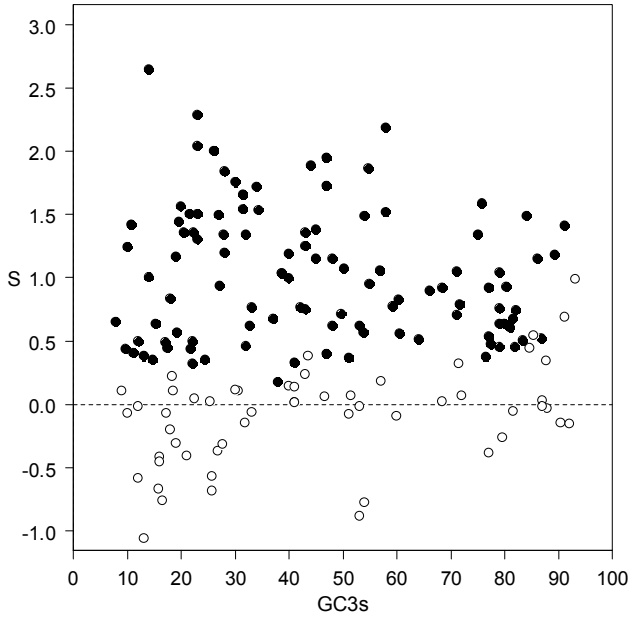

Figure 4-1 Selected codon usage bias 'S' and genomic G+C content at synonymously variable third position sites for 160 bacterial genomes.
Genomes without a significant level of selected codon usage bias are represented by open circles, filled circles show genomes where the strength of selection was significant.

value with many of the species having either S-values around or below zero with 0.09 being the average S-value for this clade. Further expansion of the Rickettsiales clade of Alpha Proteobacterial species gave similar results with the newly added species also not showing signs of significant selection and indeed showing an average S-value of -0.41. The remaining Alpha Proteobacteria show much more evidence of selected codon usage bias with 0.63 (0.87 for original 80 genome dataset) being the average S-value without the Rickettsiales. The Gamma Proteobacterial species showed similar positive selection with an average S-value of 0.77 (0.78 for the original 80 genomes) whilst the Delta and Epsilon Proteobacterial species showed, on average, weak selection with S-values of 0.30 and 0.27 respectively (0.40 for original dataset). However, it was the Firmicute and Actinobacterial species that demonstrated the most selected codon usage bias with S-value averages of 1.06 (1.19 for original dataset) and 0.99 (0.90 for original dataset) respectively; much higher than the average 0.64 S-value for the dataset as a whole. Increasing the size of the dataset appeared only to increase the resolution and did not significantly affect the overall trends found in the original study.

4.3.3 Production of the bacterial phylogeny

The first and preferred method for constructing a suitable phylogeny was to use 16S rRNA sequences. These sequences were extracted from the RDP II database (http://rdp.cme.msu.edu/) (Olsen *et al.*, 1991) and any additional sequences were obtained from the TIGR CMR (http://cmr.tigr.org). The tree that resulted looked broadly sensible when compared to other bacterial phylogenies (Olsen *et al.*, 1994; Sharp *et al.*, 2005) (Figure 4-2). However, the tree proved not to be satisfactory with the main reason being problems with the clustering of the Firmicutes. The Mollicutes subgroup of the Firmicutes did not group with the rest of the Firmicutes with the Actinobacteria unexpectedly clustering more closely instead. Imposing more restrictions on the tree proved unfeasible as not all restrictions seemed to be enforced by *MrBayes*; a known feature of the current *MrBayes* release (Ronquist & Huelsenbeck, 2003). The main problem with using 16S rRNA with such a large and varied dataset was that by the time the sequences were aligned and gaps removed the positions that were left

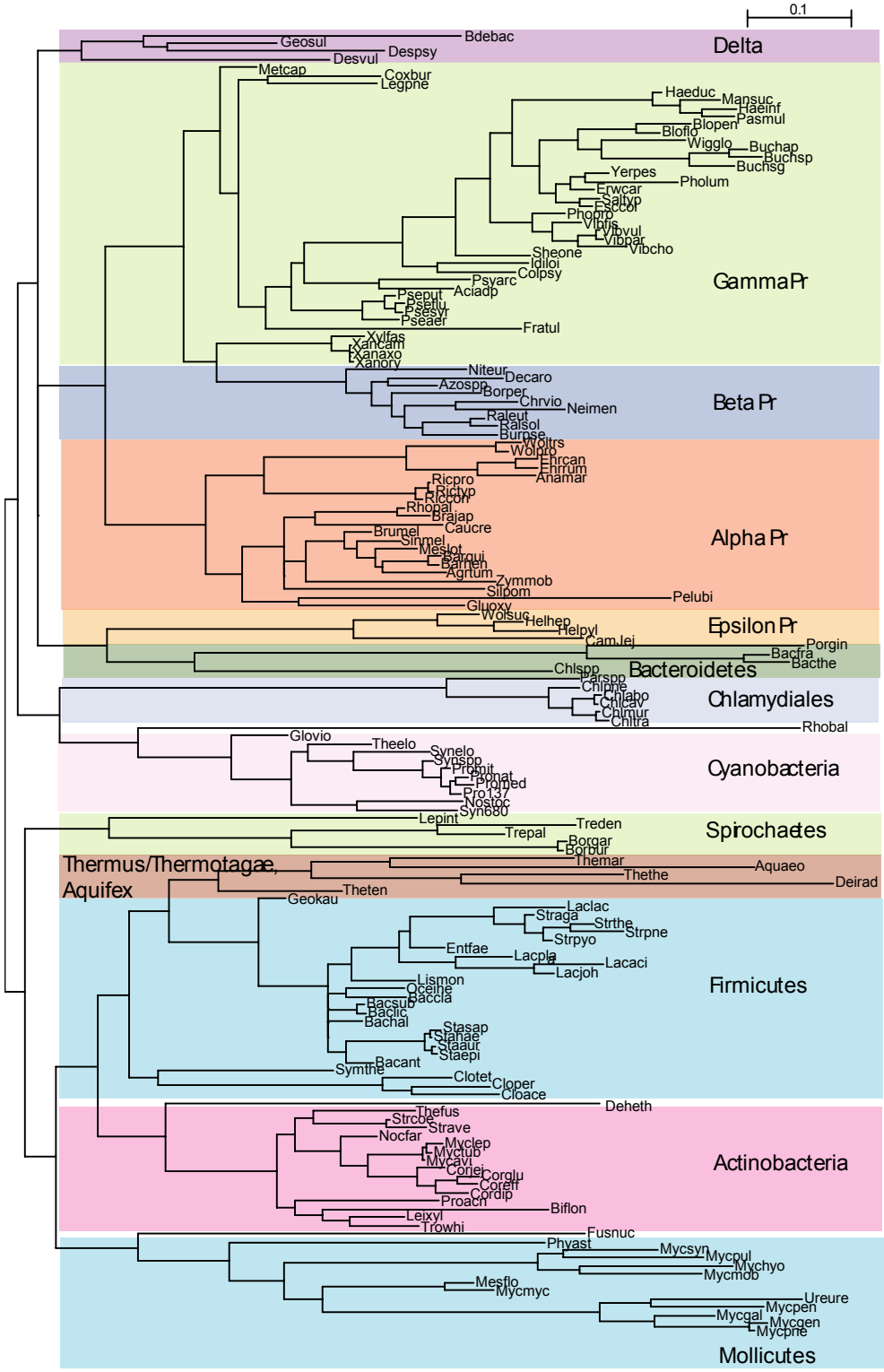

Figure 4-2 Phylogeny of the 160 completely sequenced bacterial genomes produced using 16S rRNA sequence data.
Pr signifies proteobacteria.

were largely highly conserved such that little phylogenetically informative signal remained.

An alternative solution was to use a similar method to the previous study (Sharp *et al.*, 2005) and take the ribosomal protein genes *rplA-C*, *rplB-C* and elongation factor *Tu* and use their concatenated protein sequences to draw the tree. This method proved more successful with many less multifurcations. However, near the base of the tree there were multifurcations indicating just how difficult it is to produce an accurate bacterial phylogeny. For the purposes of this research a comprehensively resolved tree was not required, nor is it likely one could have been easily achieved, and so the protein tree was kept as the best phylogeny available (Figure 4-3). It separated most of the major clades distinctly with no such problem with the Firmicutes as was found in the 16S rRNA tree. This phylogeny was, of course, very similar to the phylogeny containing only 80 genomes (Figure 4-4). The Proteobacterial relationships look very similar, only with more resolution. This could be seen in the Alpha Proteobacteria where two distinct clades of Alpha Proteobacterial were now visible, with one clade containing only the Rickettsiales. Some of the resolution at the base of the tree changed too with the Bacteroidetes not clustering with the Actinobacteria in the 160 genome phylogeny, but instead showing a close relationship to the Spirochaete and Chlamydiales clades.

4.3.4 Correlations between the number of rRNA operons, tRNA gene abundances and the strength of selection

The initial 80 genome study (Sharp *et al.*, 2005) found that the strength of selected codon usage bias seemed to be related to the degree to which speed and efficiency of growth and replication were important to the organism. The initial study compared S-values to the number of rRNA operons and tRNA abundances in each genome. This initial dataset reported a strong correlation between rRNA operon number and the strength of selected codon usage bias. The rRNA operon number was used as a surrogate for the generation time of the organism as it had been

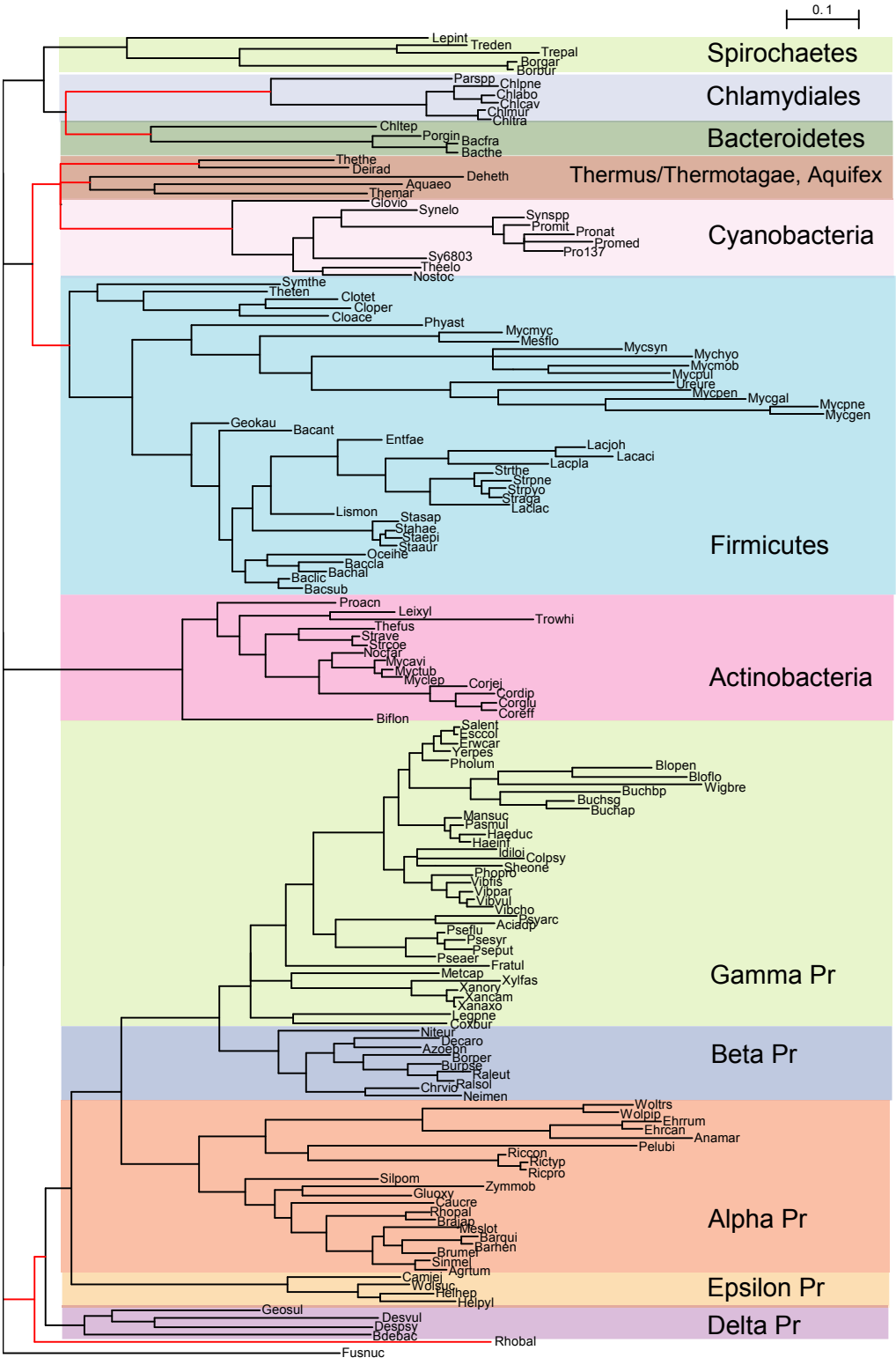

Figure 4-3 Phylogeny of the 160 genome dataset produced using ribosomal protein genes *rplA-C, rpsB-C* and EF-*Tu*.
Branches with posterior probabilities lower than 70 are marked in red, those lower than 60 were collapsed. Pr signifies Proteobacteria.

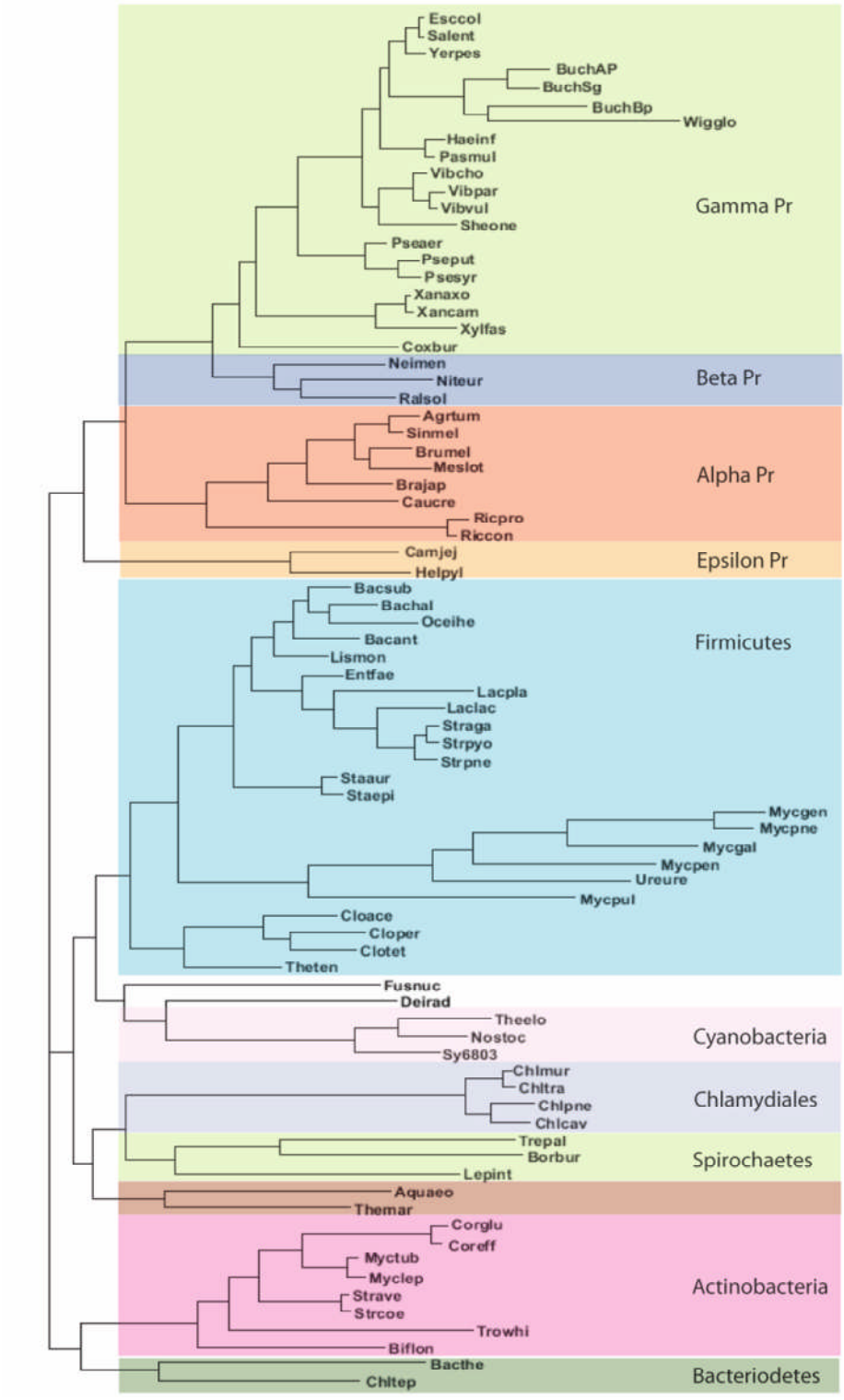

Figure 4-4 Phylogeny of the original 80 completely sequenced bacterial genomes
Produced using ribosomal protein genes rplA-C, rpsB-C and EF-Tu. Modified from Sharp *et al* 2005.

previously shown that bacterial growth rate was correlated with rRNA operon number (Klappenbach *et al.*, 2000). In this new extension of the study the larger 160 genome dataset maintained the same strong correlation (Figure 4-5).

In addition to rRNA operon number, tRNA gene copy number was also investigated using the 80 genome dataset. The abundance of different tRNAs had previously been shown to be correlated with, and apparently largely determined by, gene copy number (Kanaya *et al.*, 1999). It was noted in the original 80 genome analysis that increases in gene copy number for particular tRNA species were again correlated with an increase in selected codon usage bias in order to help to optimize translational efficiency (Sharp *et al.*, 2005). This same correlation held with the 160 genome dataset analyzed here and therefore appears to be a well established phenomenon (Figure 4-6).

In the 160 genome dataset rRNA operon numbers vary from 1 to 15 and tRNA gene copy numbers vary from 28 to 168 (Table 4-1). In addition to this, the work presented here showed a strong correlation between an increase in rRNA operon number and tRNA gene copy number (Figure 4-7) as has been shown to be the case previously (Sharp *et al.*, 2005). The previous study used rRNA operon numbers as a surrogate for an organism's growth rate but in this study of the 160 bacterial genomes generation times were obtained for 105 of the 160 bacterial genomes (Rocha, 2004).. It can be seen that the approximations were largely justified from the plots of rRNA operon number against generation time (Figure 4-8, Figure 4-9) but a more direct method was, of course, to use the newly obtained generation time data.

4.3.5 Correlations between strength of selection and generation time

When the strength of selected codon usage bias was plotted against the generation time data a strong correlation was instantly visible (Figure 4-10). Genomes with a short generation time indeed showed strong

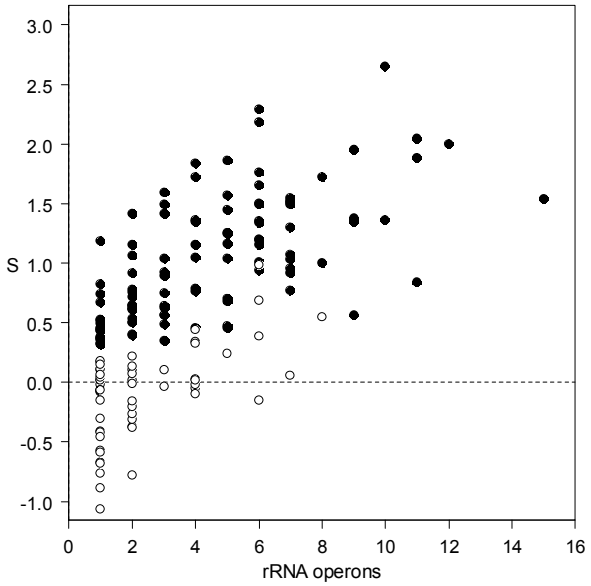

Figure 4-5 Relationship between the strength of selected codon usage bias 'S' and rRNA operon number for 160 bacterial genomes.
Genomes without a significant level of selected codon usage bias are represented by open circles; filled circles show genomes where the strength of selection was significant.

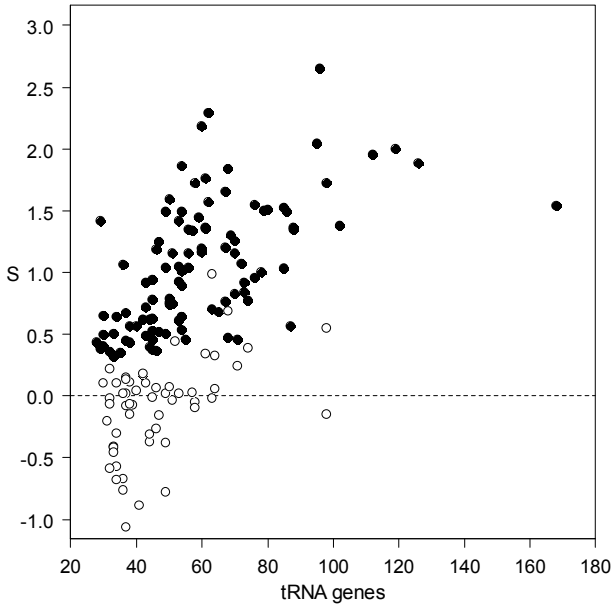

Figure 4-6 Relationship between the strength of selected codon usage bias 'S' and tRNA gene copy number for 160 bacterial genomes.
Genomes without a significant level of selected codon usage bias are represented by open circles; filled circles show genomes where the strength of selection was significant.

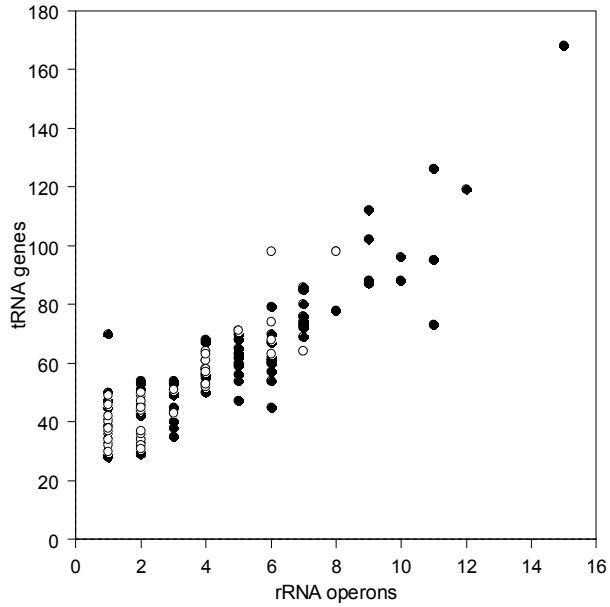

Figure 4-7 Relationship between rRNA operon number and tRNA gene copy number for 160 bacterial genomes.
Genomes without a significant level of selected codon usage bias are represented by open circles; filled circles show genomes where the strength of selection was significant.

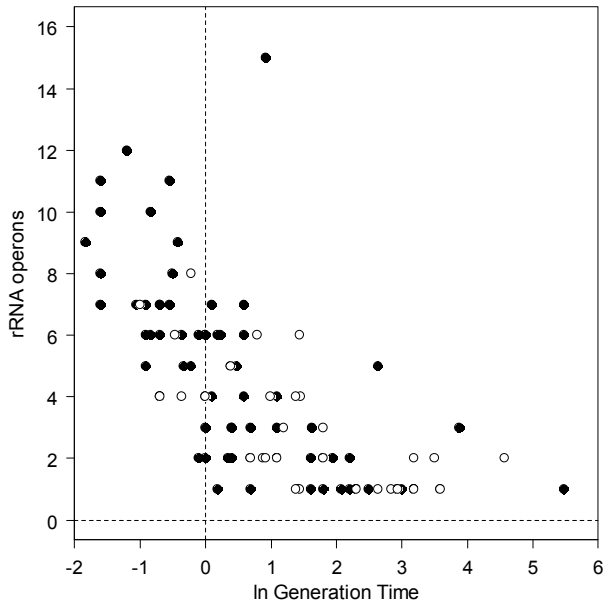

Figure 4-8 Relationship between rRNA operon number and generation time for 160 bacterial genomes.
Genomes without a significant level of selected codon usage bias are represented by open circles; filled circles show genomes where the strength of selection was significant.

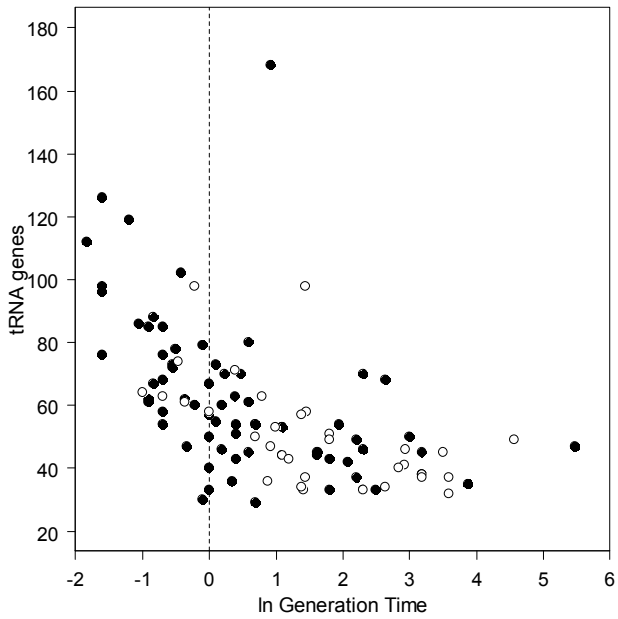

Figure 4-9 Relationship between tRNA gene copy number and generation time for 160 bacterial genomes.
Genomes without a significant level of selected codon usage bias are represented by open circles, filled circles show genomes where the strength of selection was significant.

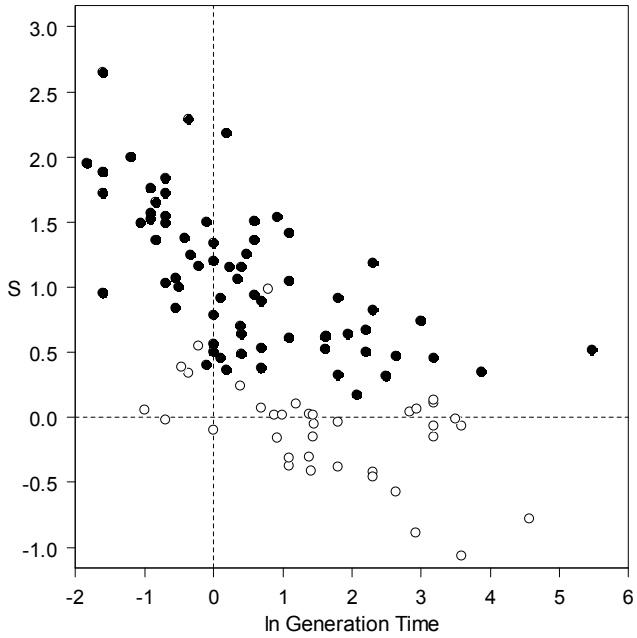

Figure 4-10 Relationship between strength of selected codon usage bias, 'S', and generation time for 102 bacterial genomes.
Genomes without a significant level of selected codon usage bias are represented by open circles, filled circles show genomes where the strength of selection was significant.

selected codon usage bias as one would expect given the hypothesis of codon usage optimization for translational efficiency. It is logical that this is the case as an organism with a fast generation time has a greater need for accurate and efficient protein production than an organism with a much slower turnover. Therefore one would expect selection to drive genes important protein production to show selected codon usage bias and use codons that pair exactly with the most abundant tRNAs in the cell. If one considers the Firmicute *Clostridium perfringens,* a bacterium with a generation time that can be as short as 8-10 minutes (Shimizu *et al.*, 2002), one can see that it also has an extremely high S-value of 2.65; the highest S-value in the entire 160 genome dataset. In contrast, organisms such as the slow growing *Mycobacterium tuberculosis* and *M. leprae* species have low S- values (0.45 and 0.52) and generation times of between 1 and 10 days.

4.3.6 Calculation of phylogeny-independent correlations

It must be noted that the correlations between rRNA operon numbers, tRNA gene copy numbers, 'S' values and generation times are all overestimated by the analysis of the data presented in the figures. This overestimation is due to the nonindependence of the data points. All 160 bacterial genomes are linked by a phylogenetic tree and one would expect closely related genomes to have similar features such as rRNA operon number, gene copy number and generation time due to their recent common ancestry. In order to derive any significance from the data presented here the effects of shared ancestry were removed using a generalized least squares (GLS) approach as implemented by Pagel in the program Continuous (Pagel, 1999). This program was used as described in the materials and methods section of this chapter. Using this method significant correlations were still achieved (Table 4-2), with the correlation between S-value and (log of) generation time being reduced from 0.654 to a still significant 0.505. This drop in correlation was quite small when compared to the drop in correlation between both rRNA operon and tRNA gene numbers and 'S' as well as the correlation between (log of) generation time and rRNA operon

Before Correction	*S value*	*rRNA.*	*tRNA*	*ORF*	*GC3s*
rRNA	**0.654**				
tRNA	**0.578**	**0.902**			
ORF	0.350	0.458	**0.587**		
GC3s	0.017	0.021	0.201	**0.634**	
ln Gen Time	**-0.652**	**-0.697**	**-0.563**	-0.264	0.095

After Correction	*S value*	*rRNA*	*tRNA*	*ORF*	*GC3s*
rRNA	0.412				
tRNA	0.334	**0.799**			
ORF	0.288	0.249	0.429		
GC3s	0.039	-0.044	0.157	0.435	
ln Gen Time	**-0.505**	-0.310	-0.141	-0.240	-0.142

Table 4-2 Correlations before and after correction for phylogenetic relatedness

number which was quite dramatic. The correlation between rRNA operon number and tRNA gene number was also maintained even after correction although all other correlations drop significantly after correction. No correlation between GC3s and S-value was found before or after correction, as was to be expected. However the most striking thing was the correlation between the strength of selected codon usage bias and (log of) generation time which held much better than the previously used surrogate for generation time, rRNA operon number.

4.4 Discussion

4.4.1 Comparing the bacterial phylogeny with other studies

The creation of an accurate bacterial phylogeny still remains somewhat of a challenge. Attempts to create a definitive bacterial phylogeny have been confounded by horizontal gene transfers and other phylogenetic noise. This problem is especially pronounced when producing an accurate deep bacterial phylogeny. The phylogeny produced in this thesis using *rplA-C*, *rpsB-C* and EF-*Tu* protein sequences (Figure 4-3) agrees broadly with recent attempts to construct prokaryotic phylogenies. The genomes of the proteobacteria show a relationship of (((Gamma, Beta), Alpha), Delta and Epsilon) which is generally well accepted. The position of the Delta and Epsilon Proteobacteria is possibly a subject of debate as some studies show epsilons do group closely with the Alpha, Beta and Gamma Proteobacteria with Deltas sitting outside the Epsilons (Ciccarelli *et al.*, 2006), whereas other studies show the Delta and Epsilon relationship to be the other way round (Bern & Goldberg, 2005; Bern *et al.*, 2006). There are relatively few genome sequences available for either clade, but it is generally supported that these two clades sit just outside the core Alpha, Beta and Gamma Proteobacterial cluster (Olsen *et al.*, 1994).

Initial analysis of the Gamma Proteobacterial clade for the 80 genome dataset showed that the tree produced (Figure 4-4) agreed largely with previously published trees (Haubold & Wiehe, 2004; Olsen *et al.*, 1994). Two main issues were discussed, firstly, it was noted that 80 genome tree

(Figure 4-4) showed *Haemophilus* to be more closely related to *Escherichia* than are *Vibrio*, in contrast to these two previous studies. The two trees produced in this thesis using 16S rRNA (Figure 4-2) and ribosomal protein genes (Figure 4-3) appeared to support the 80 genome dataset tree. Some recent attempts at the production of Gamma Proteobacterial phylogenies support this too (Battistuzzi *et al.*, 2004; Bern & Goldberg, 2005; Ciccarelli *et al.*, 2006; Wolf *et al.*, 2001). Secondly, the location of *Wigglesworthia* lying within the radiation of *Buchnera* strains was different to previous findings (Wernegreen *et al.*, 2003). Again the subsequent trees produced in this thesis using the 160 genome dataset supported the original 80 genome phylogeny with respect to this matter with *Buchnera* and *Wigglesworthia* being found together. Again recent attempts at production of an accurate phylogeny show some evidence in support of the phylogeny (Figure 4-3) presented here (Bern *et al.*, 2006; Ciccarelli *et al.*, 2006). These discrepancies in the phylogeny are relatively minor and have little consequence to any conclusions presented in this chapter. Correlations calculated changed only marginally when slightly differing tree topologies were tested as was also seen with the original dataset (Sharp *et al.*, 2005). However, it is worth noting that some features of the Proteobacterial phylogeny are yet to be conclusively proven.

If one looks at the phylogeny as a whole it can be seen that several of the groupings of major bacterial clades presented in my tree (Figure 4-3) agree with the published literature, however most attempts at bacterial phylogenies are in minor conflict with each other with reference to the positions of major clades. Recent attempts at improving the deep branch resolution may shed more light on the 160 genome tree (Figure 4-3) presented here (Bern *et al.*, 2006). The technique used was to use synapomorphies to place bacterial species in the phylogeny. A synapomorphy is a phylogenetic character that provides evidence of a shared descent. The technique used a program called *Conserv* to look for signature genes present in a clade but not in others, large insertions or deletions present only within a clade or sequence motifs well conserved within a clade but not in other clades. This method agreed in the placement of the *Buchnera/Wigglesworthia* clade with my tree and also the

placement of *Symbiobacterium thermophilum* genome with the Firmicutes which is in direct constrast to the GenBank placement of the species within the Actinobacteria. The placement of the Planctomycetes using this technique did not agree with mine but with just one species, *Rhodopirellula baltica*, being considered in my dataset this is perhaps not surprising. The Chloroflexi placement again relied on just one species in my dataset, *Dehalococcoides ethenogenes,* and also did not agree with that produced using Conserv as the 160 genome dataset showed evidence that this genome was related to species such as *Aquifex aeolicus* and not the Cyanobacteria as was suggested by Conserv. Overall the tree produced in the Bern study, like most recent studies, broadly agrees with the 160 genome ribosomal protein phylogeny presented here, Figure 4-3. It is therefore reasonable to use the phylogeny presented here for further analysis in the next chapter. All recent attempts to produce such a phylogeny show inconsistencies at the base, but they are not significant for the purposes of the research presented in this chapter as correlations corrected to take into account the effect of phylogenetic relationships between species were unaffected by small changes in the phylogeny used as was the case in the previous 80 genome study (Sharp *et al.*, 2005).

4.4.2 Comparing *S* to previous work and analyzing factors affecting the estimation of the strength of selected codon usage bias?

As stressed in the introduction to this chapter, the majority of the work concerning codon usage in bacterial genomes has involved codon usage analyses of individual bacterial species. The previous 80 genome analysis went into detail comparing the results of these analyses with the estimation of selected codon usage bias*,* 'S' (Sharp *et al.*, 2005). The broad conclusion was that 'S' did generally agree with the work discussed in individual analyses with high S-values for species where strong selected bias was previously reported, low S-values where little or no selected bias was found and even moderate S-values where weak selected bias was found. However, three factors were identified that could potentially have an effect on the calculation of an S-value. These factors arise from the method used to calculate 'S' using the four amino acids.

Firstly, the use of just four amino acids (Phe, Tyr, Ile and Asn) meant that sometimes selection could be missed if these amino acids were not subject to strong selection but other amino acids were. This was the case in the *Pseudomonas aeruginosa* genome where Ser, Ala, Thr, Arg and Gly were the main amino acids involved in selection (Grocock & Sharp, 2002). Secondly as the four amino acids only have a choice between WWC and WWU synonyms, strand bias can influence estimates of codon usage bias. This is due to the leading strand often being more G+T rich than the lagging strand combined with the fact that the majority of the 40 highly expressed genes used in the calculation of 'S' are often located on the leading strand, leading to strong strand bias giving a low S-value. This could be seen in genomes such as *Chlamydia trachomatis* where weak selection had previously been reported along with a strong influence of strand bias (Romero *et al.*, 2000), leaving the S-value calculated very low at 0.13. However when only leading strand genes were used in the calculation, to remove the effect of strand bias, the S-value calculated increased to 0.42, which was indeed indicative of weak selection. Strand bias, similarly, was shown to produce negative S-values, usually in genomes where selection was not present even when the strand bias was accounted for, such as *X. fastidiosa* and *Buchnera aphidicola* strain Bp (Sharp *et al.*, 2005). Thirdly, islands of unusual base composition can affect the estimation of 'S', if the ribosomal protein genes are located within such a region. This was the case in the Beta Proteobacterium *Nitrosomonas europaea*, which had a heavily negative S-value of -0.88. Such a negative S-value appeared to be a result of such unusual base composition.

Since the 80 genome analysis was done, several new papers have been published looking at newly acquired bacterial genomes in the 160 genome dataset. The genome of *Deinococcus radiodurans* has been analyzed (Liu, 2006) and the authors indicate that translational selection is the main factor influencing codon usage in the genome. This is in agreement with an S-value of 1.49, which is indeed strong. The value is the same as the 'S' value estimated for *E. coli*, which has been the archetypal example of a

species with strongly selected codon usage bias. We showed evidence of translational selection in the *Shewanella oneidensis* genome (Henry & Sharp, 2007), see appendix B, which agrees with the S-value of 1.38, shown here. The previous chapter also found evidence for translational selection in *Bdellovibrio bacteriovorus* which again has a moderately high S-value of 1.06.

A study looking at the *Bartonella* species has also been published (Das *et al.*, 2005) argued that the primary factor influencing codon usage in these bacteria is strand bias and that this is a sign of selection along with the majority of genes being located on the leading strand; why this is evidence for selection is unclear and indeed the S-value for *Bartonella quintana* in this thesis was just -0.315 whilst that for *Bartonella henselae* was also low at -0.373. The authors performed a correspondence analysis on the data and showed that strand bias is present but no evidence was shown that translational selection was responsible for any of the remaining axes. In the paper they took the *Bartonella quintana* origin of replication to be the region of the genome between BQ13580 and BQ00010, and the terminus to be around 723kb from the origin, which is the *metS* gene BQ06090. To try to find evidence of translational selection, I used the same positions and assigned genes as either lagging or leading on this basis (excluding 10 genes either side of the terminus). This gave 708 leading strand genes with 34 ribosomal genes on leading strand. A calculation to estimate the strength of selected codon usage bias only using these leading strand genes gave an S-value of -0.104 which indicates no significant degree of selection. Analysis of the leading strand genes did show that even within the leading strand there was variation in C3s and U3s. A region of the genome encompassing approximately 200 leading strand genes seemed to have a stronger U preference and included the ribosomal genes but strong alterations in codon usage were not really evident. One can be reasonably confident, therefore, that there is no strong evidence of selected codon usage bias in this genome.

It can be noticed that some of the genomes in the dataset have negative S-values. It may be expected that the minimal S-value would be around zero and illustrate a case where there was no selection present. However, the table shows several S-values that are lower than zero. As mentioned previously in this section this is often due to the effects of strand bias or islands of unusual bas composition within the genome. To get around this the S-value could be calculated using genes solely on the leading strand. However, it is often difficult to locate origins and termini of replication. Also gene location is often not very well conserved even among closely related species. From the 160 genome dataset only 13 genomes show a significantly negative strength of selected codon usage bias, that is to say they lie below the 95% confidence interval. Several of the major groups of bacteria appear to be affected by this, most notably the Beta and Alpha Proteobacteria. The Alpha Proteobacteria have eight species with significantly negative S-values whilst the Beta Proteobacteria, although with only one significantly negative S-value, contains genomes with probable underestimates of 'S' leaving only one *Ralstonia* species with a significantly positive S-value. Of the 13 significantly negative S-values, eight are newly reported here whilst the other five were part of the original study containing just 80 of the genomes.

Included in the 13 genomes with a significantly negative predicted strength of selected codon usage bias were seven genomes of the order Rickettsiales. The Rickettsiales stood out as having an unusually high number of genomes with negative S-values. Correspondence analysis on the two *Ehrlichia* genomes showed that the primary factor influencing codon usage in these species was a strong strand bias with a 0.94 correlation between the primary correspondence analysis axis and GT3s in both *Ehrlichia canis* and *Ehrlichia ruminatium*. A similar trend was shown by *Anaplasma marginale* with the primary correspondence analysis axis showing a Pearson correlation of 0.78 with GT3s. A simple correspondence analysis on RSCU data was, however, not able to identify significant trends in codon usage bias in the two *Rickettsia* and *Wolbachia* strains. It appeared that rare codons such as arginine were the main causes of this problem and so a within-block correspondence analysis was performed on

the *Rickettsia conorii* genome. This method eliminated the arginine related axes and revealed a primary axis related to GC3s and a secondary axis highly correlated with K3s. Although strand bias seems to be the main cause of the large negative S-values seen, there was no indication that any of the Rickettsiales showed any influence of translational selection. Indeed early codon usage analysis of *Rickettsia prowazekii* (Andersson & Sharp, 1996) also reported no translational selection. The *Ehrlichia*, *Wolbachia*, *Rickettsia* and *Anaplasma* genomes in the dataset showed very little evidence of selection with only one ribosomal operon in each genome and relatively few tRNAs with numbers ranging from 33-36 tRNA species, which would be associated with a genome with little selected codon usage bias. This demonstrates that although it can be argued that the strong strand bias in these genomes may be artificially lowering the calculated strength of selected codon usage bias, other features of the genome show that selection is very unlikely to be influencing codon usage in these species.

The genome with the largest negative S-value is *Blochmannia floridanus*, with a value of -1.067. When a correspondence analysis was done on the genome it could be seen that the major factor influencing codon usage in this genome was strand bias (correlation with axis one: 0.92). The leading strand was G+T rich (GT3s 0.60) compared to the lagging strand (GT3s 0.44). When the highly expressed genes were compared to those just on the leading strand the S-value rose to around -0.30. It is once again probable that although the strong strand bias has reduced the 'S' estimates there is little evidence for any selection influencing codon usage bias in this genome. The genome contains only one rRNA operon, just 37 tRNA genes and also has a long generation time of 36 hours suggesting that little selection is at work here.

The use of 'S' does, therefore, seem to be largely accurate and in agreement with published work on individual genomes, with some exceptions due to the peculiarities of the 'S' statistic. However, as long as one is aware of such exceptions this measure of selection can be used reliably.

4.4.3 Explaining variation in selected codon usage bias

Throughout this chapter it has been suggested that a high S-value indicates that an organism has a competitive lifestyle where a fast generation time is advantageous whilst organisms with a low value of 'S' have a life history where exponential growth was unimportant. There is, of course, an alternative to this where the lack of selected codon usage bias may be due to the greater impact of genetic drift resulting from a population structure with a low long-term effective population size. This may be combined with interference between linked synonymous sites due to a lack of recombination. However, it is hard to know the long-term evolutionary effective population size relevant to codon usage. It does seem possible that the life histories of some of the bacteria analyzed may lead to a low effective population size. In the previous 80 genome dataset it was reported that many of the genomes with low strength of selected codon usage bias were intracellular parasites. The new species added to create the 160 genome dataset also appear to support this trend. The new species added to the Rickettsilaes clade are largely parasitic and have low selected codon usage bias. The *Bartonella* species are also parasitic in nature and show similarly low selected codon usage bias. As stated in the original 80 genome analysis it is very difficult to separate the effects of low population size and growth rate from the environmental factors that promote fast exponential growth but both seem to be important in influencing codon usage in bacterial species.

It has been shown in this chapter that the strength of selected codon usage bias can vary greatly between different bacterial species. The doubling of the dataset from 80 to 160 genomes has left the conclusions of the original 80 genome analysis (Sharp *et al.*, 2005) unchanged and even better supported. Additionally, the strength of selected codon usage bias has been shown to be strongly associated with the lifestyle of an organism, with bacteria with rapid growth showing strong selected codon usage bias correlated with fast generation times. This correlation was still significant even after correction to take into account phylogenetic relationships.

Chapter 5:
Exploring switches in optimal codons

5.1 Introduction

In addition to variation in strength of selection, as discussed in the previous chapter, bacterial genomes also show variation in the direction of selection. Indeed, two organisms although both displaying a strong selected codon usage bias may have preference for totally different synonymous codons. It has long been clear that codon usage preference is not always consistent between genomes (Grantham *et al.*, 1980b). This leads to perhaps the most intriguing question in the analysis of codon usage in bacterial genomes, just how does this switch in codon preference occur?

If codon usage and tRNA abundances co-evolved, and are well co-adapted, then a shift in either must be detrimental to the organism's well being. A change in codon usage would mean that codon usage would no longer be as well adapted to the tRNAs present in the cell. Alternatively, a change in tRNAs, either abundance or anticodons used, would also result in a less well adapted relationship between codon usage and tRNAs present. It is hard to imagine that both codon and tRNA usage could change at once, but the mechanism by which codon and tRNA preference can change is not well understood. A mechanism has been put forward in order to explain how codon and tRNA usage could change without a relaxation of selection. Such a theory was put forward by Shields as discussed in the introduction to this thesis.

This chapter looks at how selected codon usage bias changes across the 160 genome bacterial dataset and investigates the factors that seem to shape selected codon usage bias across these species.

5.2 Materials and Methods

5.2.1 Modifying the 'S' statistic for to look at two fold degenerate codon switching

The case of two fold degenerate codons is the simplest to look at and it was, therefore, the first case examined. To look at this change in codon preference a method was devised to measure selection for particular codons by modifying the 'S' statistic. This modified S-value is calculated in exactly the same way the original S-value (S_{WWY}), but instead of taking the codon usage of phenylalanine, tyrosine, isoleucine and asparagine together one just considers the amino acid of interest. This S_{aa} value can then be plotted against the S-value calculated using the four amino acids with codons of the form WWY to look for switches in codon preference. When the S_{WWY} was calculated it was known that, for each of the four amino acids considered, the C-ending codon was always preferred over the U-ending codon. When considering the 'S_{aa}' value for each amino acid it is not important which codon was selected as the preferred (**C1**) codon, as a switch in preference was all that was being examined. Consistency was maintained to ensure that all the resultant switching plots could be interpreted in the same manner. In the analysis of two fold degenerate codons the U or A ending codon was always considered **C1** and the G or C ending codon was always considered **C2.** Codon switching plots of S_{aa} vs S_{WWY} were then produced for each of the two fold degenerate amino acids.

5.2.2 Modifying the 'S' statistic for to look at four fold degenerate codon switching

Plotting the codon usage of four fold degenerate codons presented more of a problem. One solution to this problem was to group NNA+NNU codons together and NNG+NNC codons together and plot them in a similar fashion to the two fold degenerate codon plots. However, this masks much of the information that can be obtained by examining switches in selected codon usage bias. To solve this dilemma a new way of graphing the data was needed so that no information was lost.

To solve this problem a pseudo-3-dimensional plot was created where the x-axis was calculated by grouping NNA+NNG and NNU+NNC codons together, with A+G ending codons as the nominally preferred, **C1,** codons. The y-axis was created by grouping NNA+NNU and NNG+NNC codons with A+U ending codons as **C1** codons. This means that each corner effectively represented one of the four possible codons with NNU, NNA, NNG and NNC codons respectively going clockwise from top left to bottom left. The radius of each data point was directly proportional to the strength of selected codon usage bias using the S_{WWY} value, thus creating a pseudo-z-axis. These plots do not mask any codon preference information and so were adopted as the preferred means of representing codon preference for four-fold degenerate codons.

5.3 Results

5.3.1 A graphical method to explore two-fold degenerate codon preference switching

In order to view codon switches across all 160 genomes in the dataset a graphical method was employed plotting the strength of selected codon usage bias, calculated using the four codons of the form WWY (Phe, Tyr, Ile, Asn), on the x-axis and the strength of selection for each individual amino acid of interest plotted on the y-axis calculated as described in this chapter's materials and methods. Such a method provided any easy way of spotting general trends in codon usage preference across the huge 160 genome dataset. The raw data used to plot the graphs presented in this chapter can all be found in appendix C.

The plots for two fold degenerate amino acids with synonymous codons of the form NNY gave results that were largely expected. The phenylalanine, tyrosine, isoleucine and asparagine plots (Figure 5-1) all showed that WWC was preferred over WWU for each of the amino acids. The way 'S' is defined means that this should be the case as S_{WWY} is calculated using these four amino acids, these graphs are shown here more as a matter of

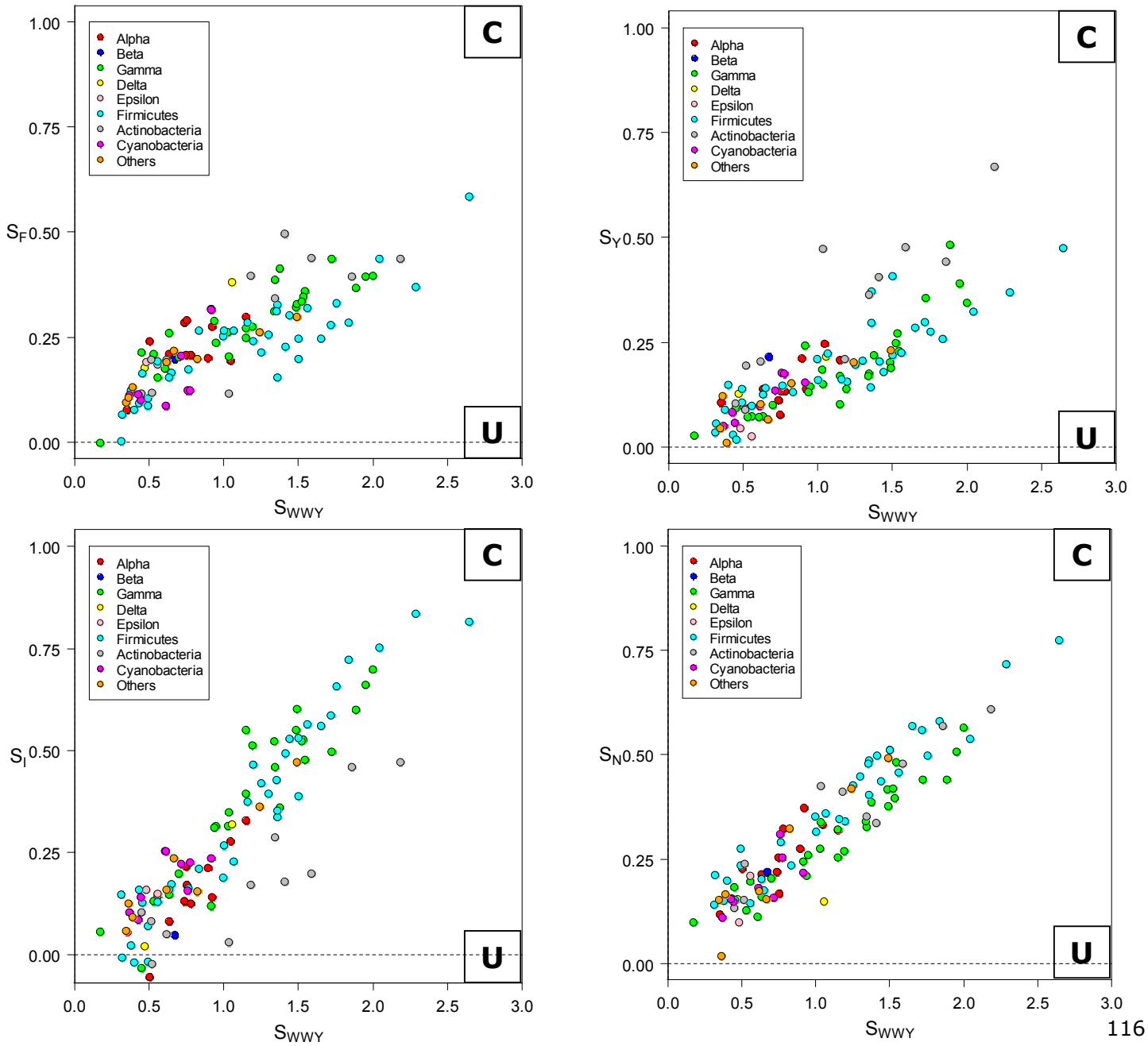

Figure 5-1 A plot showing the switching of selected codon usage bias for the amino acids phenylalanine, tyrosine, isoleucine and asparagine.

completeness. More interesting were those plots for the other two fold degenerates of the form NNY. The aspargine and histidine amino acids (see Figure 5-2 and Figure 5-3) both showed preference for NNC over NNU the selection for NNC appeared to be just as clear as for the four amino acids used in the calculation of S_{WWY}. The cysteine amino acid is rare, especially in the highly expressed dataset, and so S-values for this amino acid are not always reliable as they are based on few codons. Therefore, it appears that for amino acids with synonymous codons of the form NNY the NNC codon is always preferred over NNU when the genome is subject to translational selection. This meant that for this 160 genome dataset, covering a wide cross section of bacterial species, no switching of optimal codon usage was evident for NNY codons.

In contrast, when two fold degenerate codons of the form NNR were considered there appeared to be evidence of a switch in optimal codon usage. The first amino acid to be considered was the amino acid glutamine where a plot of S_Q against S_{WWY} (Figure 5-4) indicated clear switching of selected codon usage bias. The switching did not appear to occur randomly but instead involved species within major groups of bacteria. The most notable changes were observed in the four groups best represented within the dataset; the Firmicutes, Actinobacteria, Alpha Proteobacteria and Gamma Proteobacteria. In bacterial species with high S_{WWY} values, a preference for CAA is shown by the Firmicutes and many of the Gamma Proteobacteria; although to a much lesser degree, with some of the Gamma Proteobacteria having a relatively neutral glutamine codon preference. In contrast, the Alpha Proteobacteria, excluding the Rickettsiales clade (which mostly shows little evidence of selection), and the Actinobacteria show preference for the CAG ending codon. In addition *Xanthomonas campestris, X. oryzae* and *X. axonopodis* of the Gamma Proteobacteria appear to show preference for the CAG ending codon. Such a switch in codon preference suggests that selective pressure switched once in a common ancestor for each of these clades and has been maintained in these major bacterial lineages. It is also possible that some smaller groups such as the Delta Proteobacteria and some Cyanobacteria may exhibit a switch in codon usage but the resolution is not high enough and more

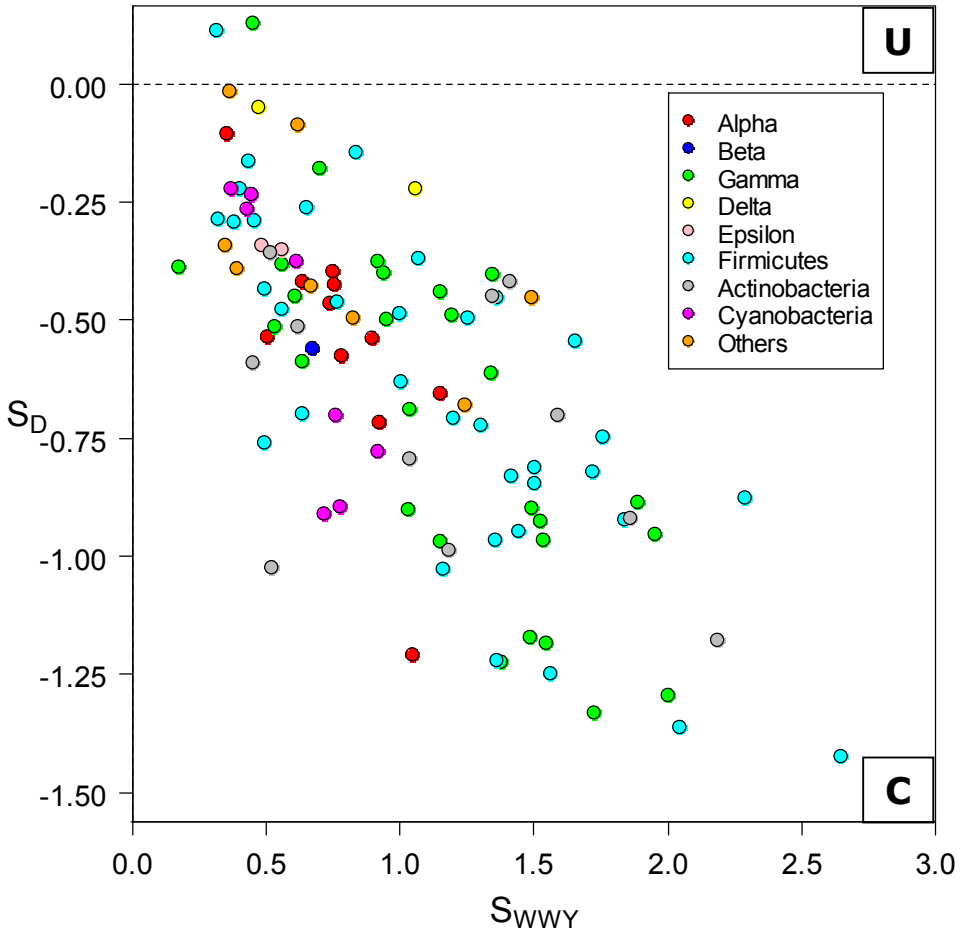

Figure 5-2 A plot showing the switching of selected codon usage bias for the amino acid aspartate.

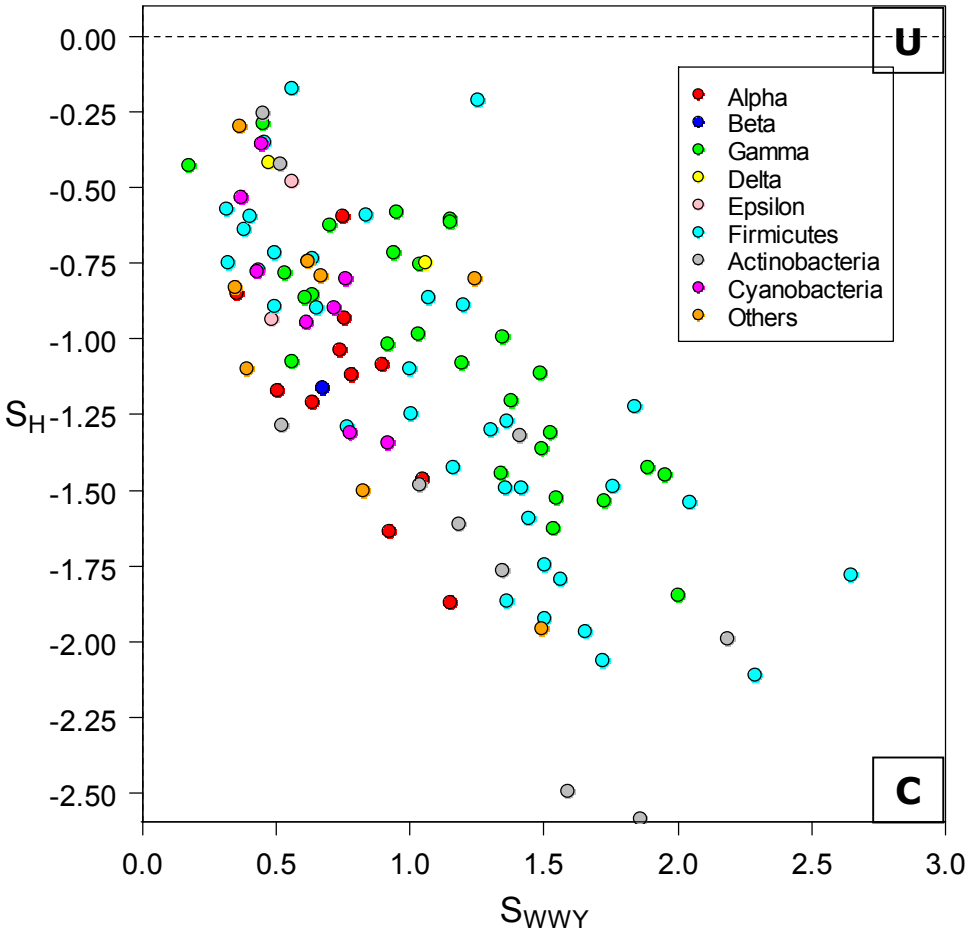

Figure 5-3 A plot showing the switching of selected codon usage bias for the amino acid histidine.

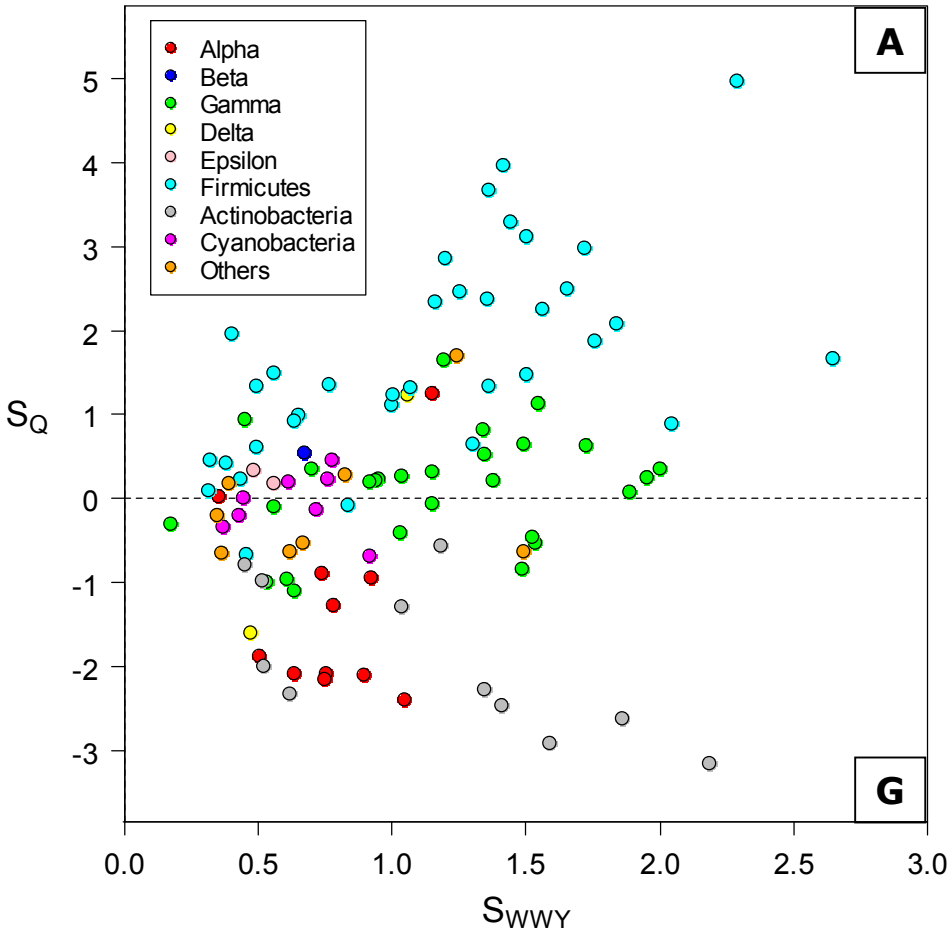

Figure 5-4 A plot showing the switching of selected codon usage bias for the amino acid glutamine

genomes would be needed to be more confident about this. The selected codon bias affecting glutamine codons is also very high in some groups with S_Q values ranging as high as 5 and as low as -3; this high degree of selection is most notable in the Firmicutes, where many genomes have S_Q values of over 2, and the Actinobacteria where many genomes have S_Q values of less than -2.

When the amino acid glutamate was considered (Figure 5-5) the plot of S_E vs S_{WWY} also indicated that optimal codon switching was taking place. The majority of bacterial species showed a preference for the GAA codon but there was a clear switch in preference among all the Actinobacteria to the GAG codon. Some genomes displayed no significant preference as was the case with many Gamma Proteobacterial genomes such as *Haemophilus influenzae* (S_Q = -0.001) and *Pasturella muticoda* (S_Q = -0.024). So again the Actinobacteria showed a preference for NNG codons but this time no other switch appeared to have taken place with the Alpha Proteobacteria (which preferred GAA). As a whole, selection for glutamate was also found to be less strong than for glutamine with S_E only ranging from 2.5 maximum GAA preference for the Firmicutes to -1.5 GAG preference for the Actinobacteria.

The final abundant two fold degenerate amino acid, lysine, was then considered. The plot of S_K vs S_{WWY} for the amino acid lysine (Figure 5-6) showed evidence for another switch in selected codon usage bias. Again the Actinobacteria showed evidence of an alternative codon preference, preferring the AAG codon, along with the Alpha Proteobacteria. The three *Lactobacillus* species in the Firmicutes also showed a preference for AAG (S_K ranging from -1.68 to -0.55) but the majority of the Firmicutes showed a relatively strong AAA preference. The codon preference of the Gamma Proteobacteria was weak and seemed to be quite neutral. In fact, for all NNR codons the Gamma Proteobacteria showed no particularly strong directional selection apart from the *Xanthomonas* species which showed a AAG preference just as the CAG codon was preferred in the *Xanthomonas* species for glutamine. For this amino acid selection for the A ending codon

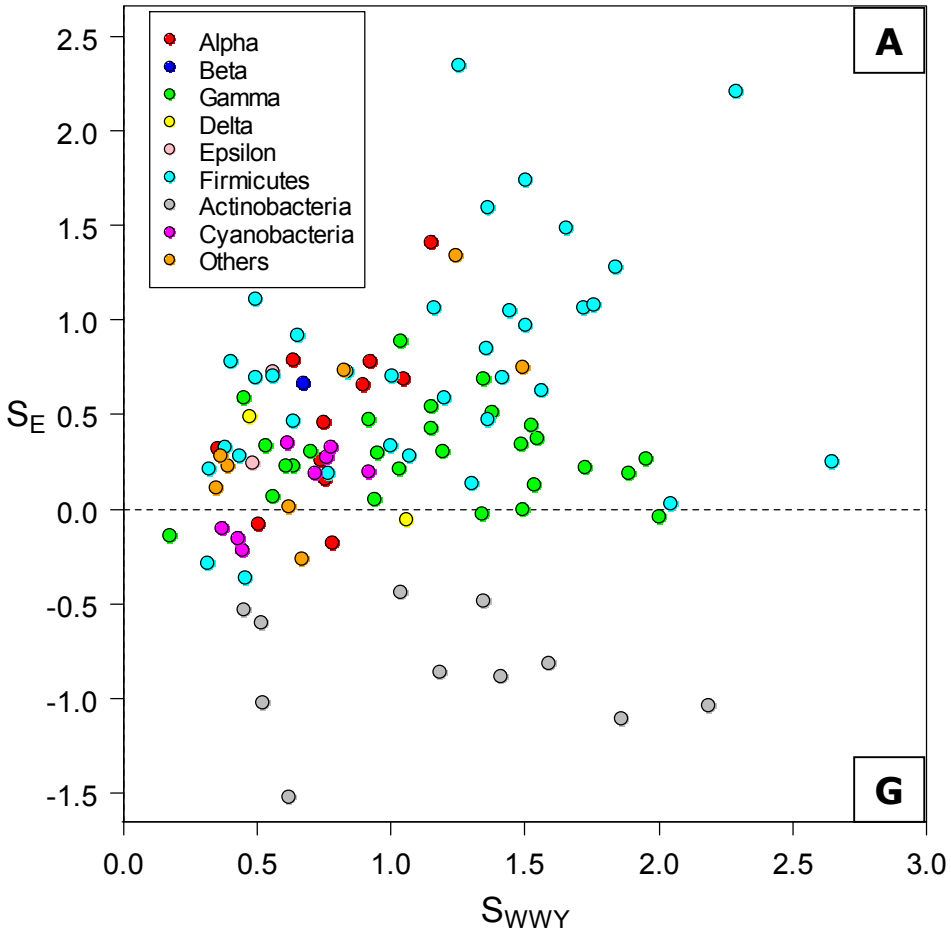

Figure 5-5 A plot showing the switching of selected codon usage bias for the amino acid glutamate.

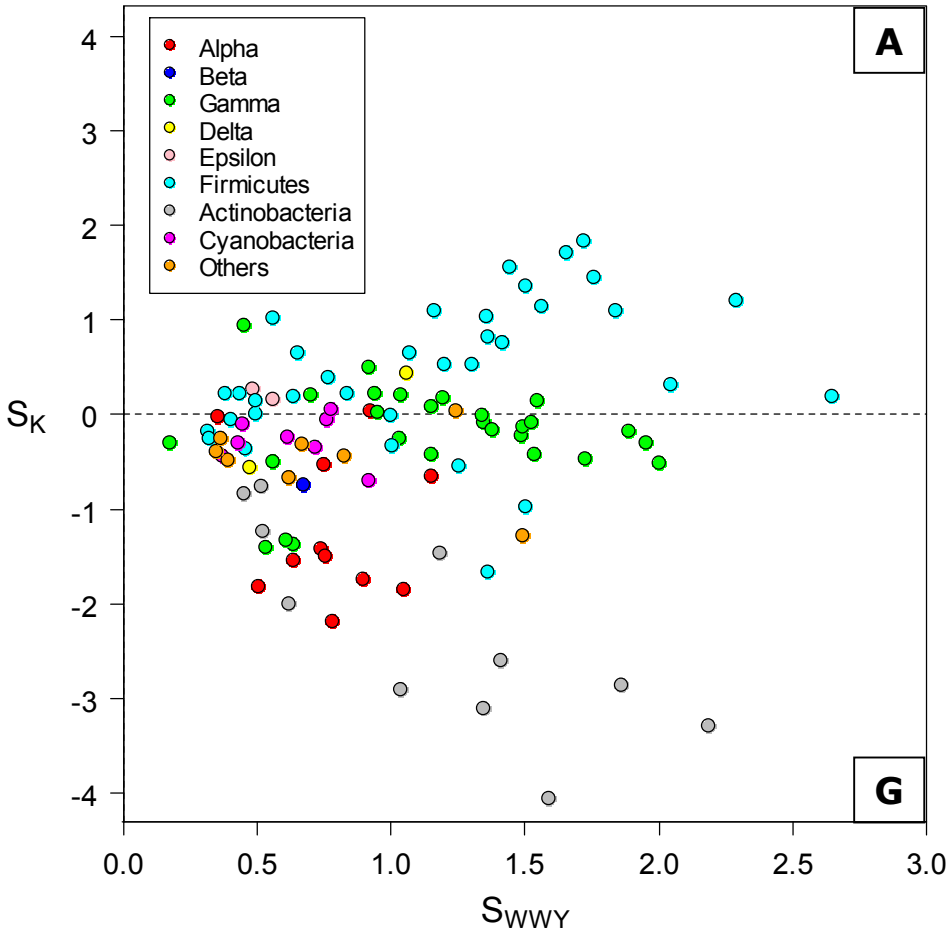

Figure 5-6 A plot showing the switching of selected codon usage bias for the amino acid lysine.

was a lot weaker than the other amino acids with codons of the form NNR. The Firmicutes were the only major group to show AAA selection with a maximum S_K of approximately 2. In contrast the selection for AAG was stronger than the other two amino acids with codons of the form NNR with a S_K of approximately -4 for some Actinobacteria.

It therefore appears that two fold degenerate amino acids can be split into two categories with amino acids using NNY codons being resistant to directional selection and maintaining a NNC over NNU preference. In contrast amino acids using NNR codons experience switches in optimal codon preference across major groups.

5.3.2 Exploring switches in codon preference for four-fold degenerate codons

Plotting four-fold degenerates was more complex: plots were constructed where each corner of the plot represented a synonymous codon and the position of each genome on the plot showed how strong the preference for each of the four codons was. The strength of selected codon usage bias, as calculated using the four amino acids with codons of the form WWY, was shown by the size of the point. The construction of these plots is explained in detail in this chapter's materials and methods section and the data plotted is available in appendix C.

Proline appeared to be the only four-fold degenerate amino acid where purine ending codons are largely preferred over the pyrimidine alternatives (see Figure 5-7). The majority of Firmicutes and Gamma Proteobacteria appear to have a preference for the A ending proline codon (CCA) along with some of the Actinobacteria, in particular the *Corneybacteria*. The *Bacillus* species were swayed towards the pyrimidine ending codon CCU as were *Lactobacillus plantarum* and *Listeria monocytogenes*. Of the species which prefer G+C ending codons the most frequently preferred was the purine ending codon CCG, with the Alpha Proteobacteria, some of the Actinobacteria and the *Xanthomonas* species (Gamma Proteobacteria) showed a CCG preference. The Enterobacteria *Salmonella enterica* and

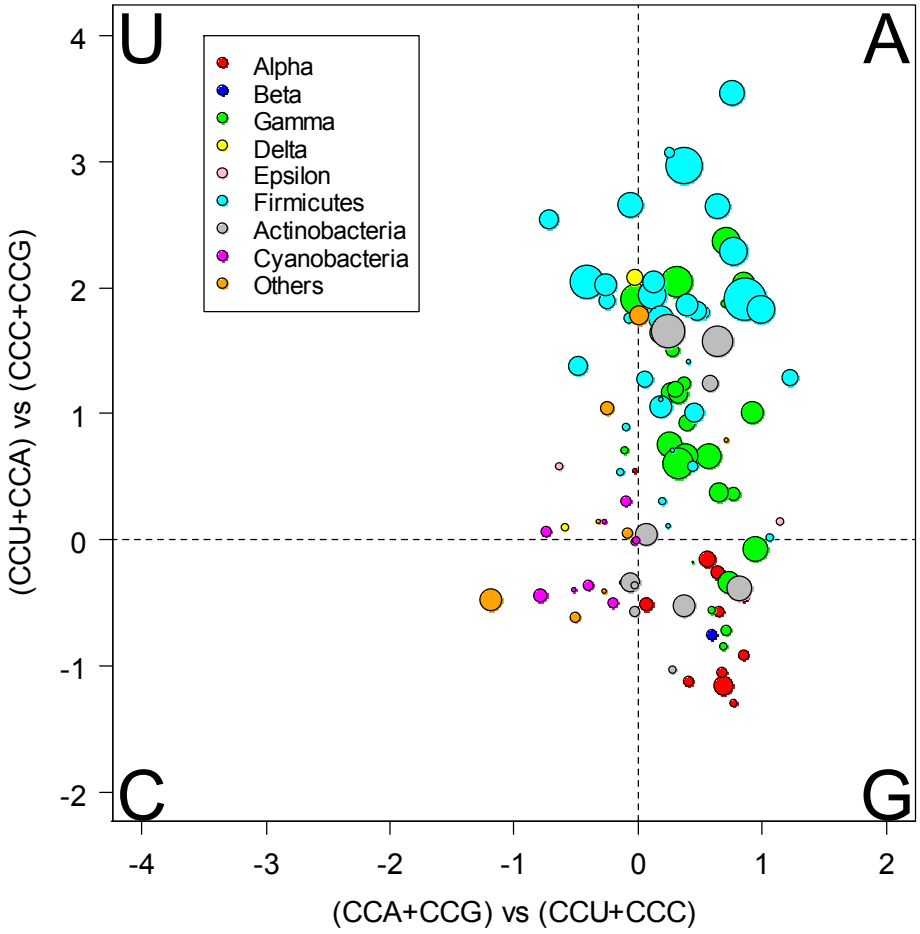

Figure 5-7 A plot showing the switching of selected codon usage bias for the amino acid proline.

Escherichia coli both showed a CCG preference unlike the remaining Gamma Proteobacteria. There were genomes that preferred the C-ending codon, they included a subgroup of the Cyanobacteria and the *Deinococcus radiodurans* genome. Threonine was more typical of the four degenerate amino acids with pyrimidine-ending codons being preferred. There was a clear distinction between those genomes preferring ACU and those preferring ACC (see Figure 5-8). The U-ending codons were preferred by the Firmicutes and most of the Gamma Proteobacteria. Two of the Clostridia species, *Clostridium acetobutylicum* and *Clostridium tetani* (Firmicutes) showed weak NNA over NNU but like other Firmicutes and the majority of Gamma Proteobacteria showed very strong NNW over NNS codon preference. In contrast the Actinobacteria, Alpha Proteobacteria and possibly the Beta Proteobacteria (although there were few genomes displaying strong evidence of selection amongst the Beta Proteobacteria so one cannot be sure) preferred ACC along with the *Xanthomonas* species from the Gamma Proteobacteria (a strong NNS over NNW preference). It also appeared that some of the Cyanobacteria and *Deinococcus radiodurans* preferred the C-ending codon, as they appear to do in the majority of the 4-fold degenerate amino acids examined here.

Codon preference for valine (Figure 5-9) was firmly in the direction of the GUU codon with little support for any other codon except for GUC in the case of *Deinococcus Radiodurans* and *Mycobacteria avium*.

The plot for alanine (Figure 5-10) also showed that most genomes have a preference for GCW over GCS and, in particular, the GCU codon as opposed to GCA. Some Gamma Proteobacteria and Firmicute genomes showed some slight GCA preference and these will be discussed later in this chapter. There is also a small indication that the Cyanobacterial genomes of *Synechococcus sp*. strain WH8102, *Prochlorococcus marinus* strain MIT9313, *Deinococcus Radiodurans* and also some Actinobacteria may again show some preference for C-ending codons. The main preferred codon, however, appears to be the codon GCU.

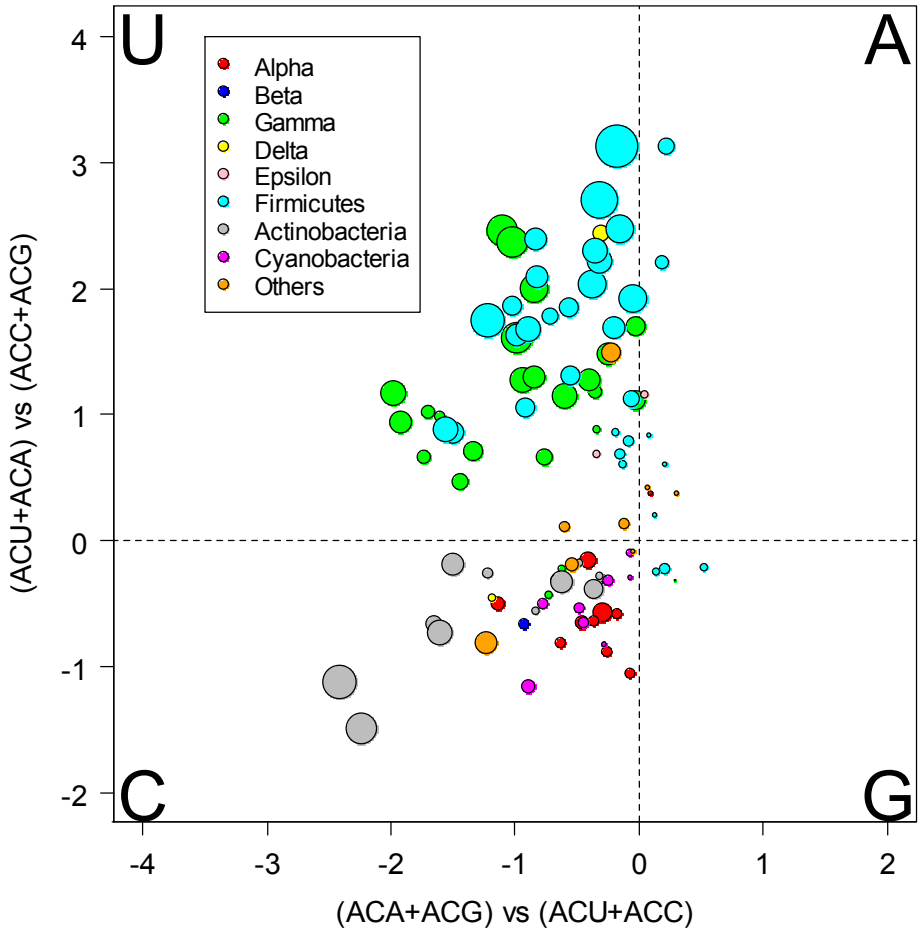

Figure 5-8 A plot showing the switching of selected codon usage bias for the amino acid threonine.

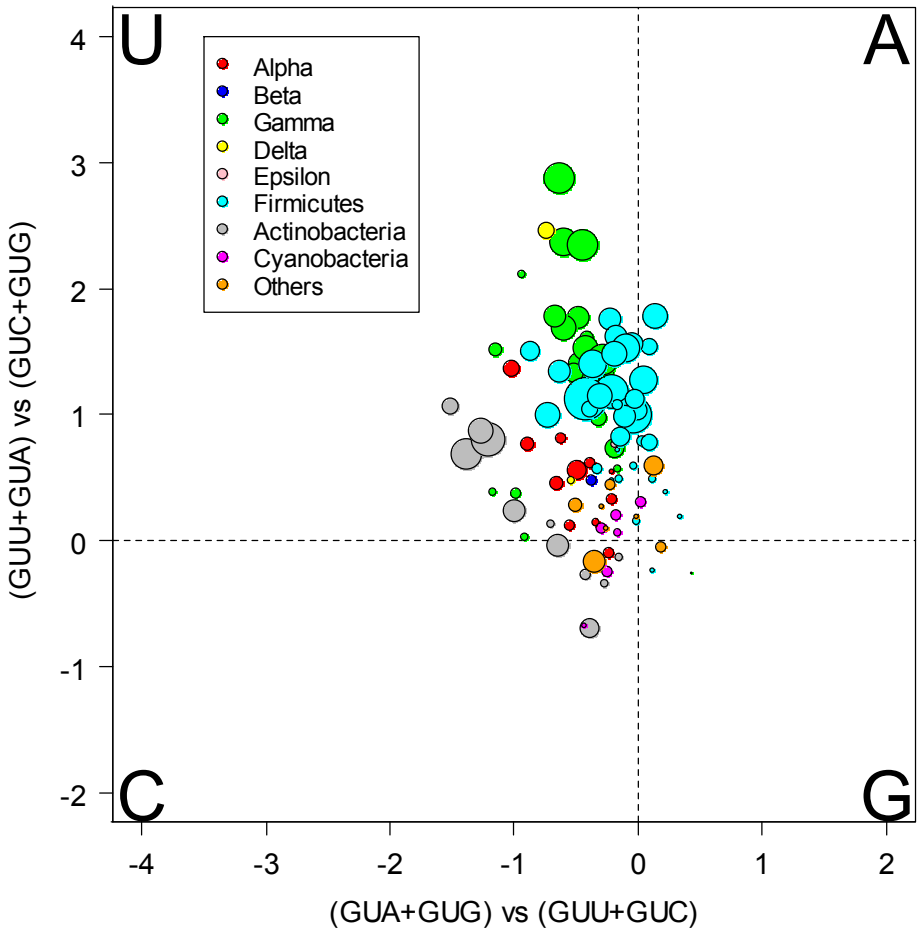

Figure 5-9 A plot showing the switching of selected codon usage bias for the amino acid valine.

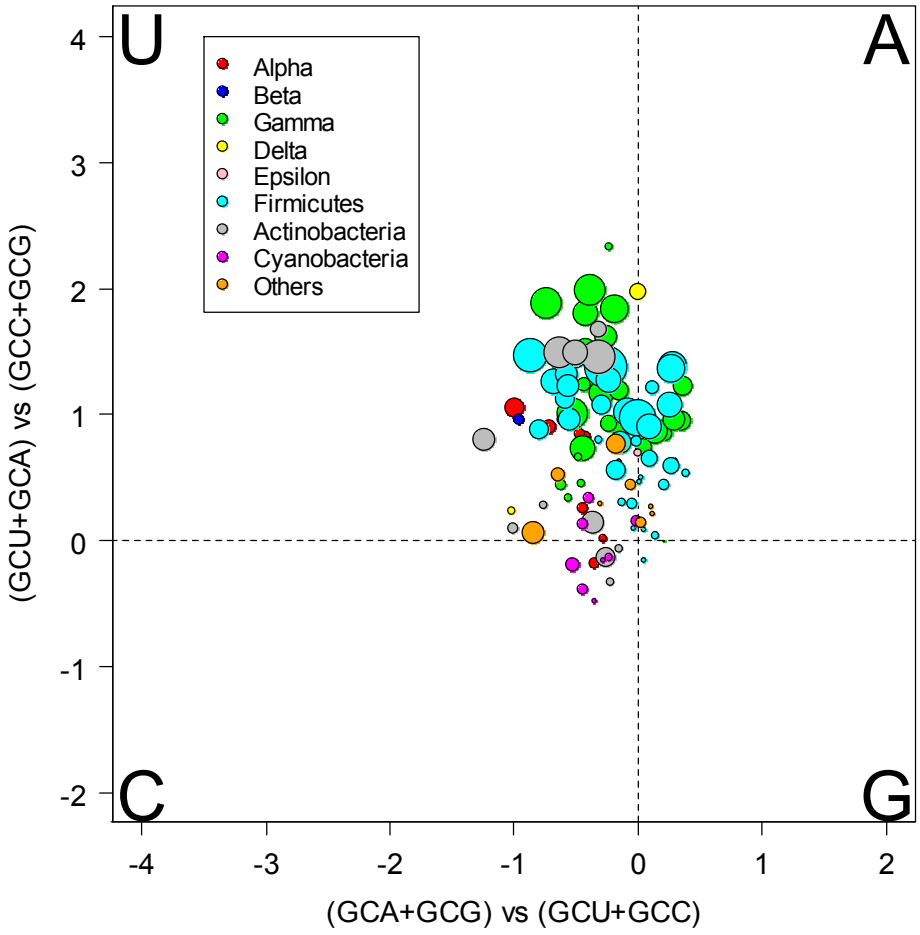

Figure 5-10 A plot showing the switching of selected codon usage bias for the amino acid alanine.

For the glycine plot (Figure 5-11) there was a strong preference for pyrimidine ending codons with GGU appearing to be the preferred codon of the two. The *Vibrio* genomes of the Gamma Proteobacteria showed a particularly strong pyrimidine preference as did *Salmonella enterica* and *Escherichia coli*. Some Actinobacterial species showed some weak GGC preference, particularly in genomes such as the C*orynebacteria*. A few Cyanobacteria also followed with some weak GGC preference in contrast to the GGU preference that was predominant over all genomes as a whole.

The analysis of the four fold degenerate amino acids was more difficult to examine but some general trends in codon usage were established. Firstly, the proline and threonine amino acids showed splits in preference largely between genomes that preferred A+U ending codons to those that favoured G+C ending codons. Bacteria seemed to largely use NNU codons for valine and alanine in most circumstances whilst for glycine there was a varying amount of selection across genomes for GGU ending codons with the Gamma Proteobacteria having the largest GGU preference.

5.3.3 Using optimal codon preference to assign significance to the general switching patterns

The switching plots showed that changes in selected codon usage bias were indeed occurring in genomes where the S-value (S_{WWY}) provided evidence of a significant level of selection. However, it was also interesting to look at codon preferences across all the genomes in the dataset to look for cases where S_{WWY} may have missed a case of selection and to analyze how useful S_{WWY} is in looking at selection. This analysis also allowed a more in depth look at selected codon bias across all genomes, and in particular six-fold degenerate codons which were not looked at using the graphical methods in the previous section. Finally, where it was suspected that a genome was subject to selection this form of analysis allowed one to say whether a particular codon was used at a significantly higher level in the highly expressed genes as compared to the genome as a whole.

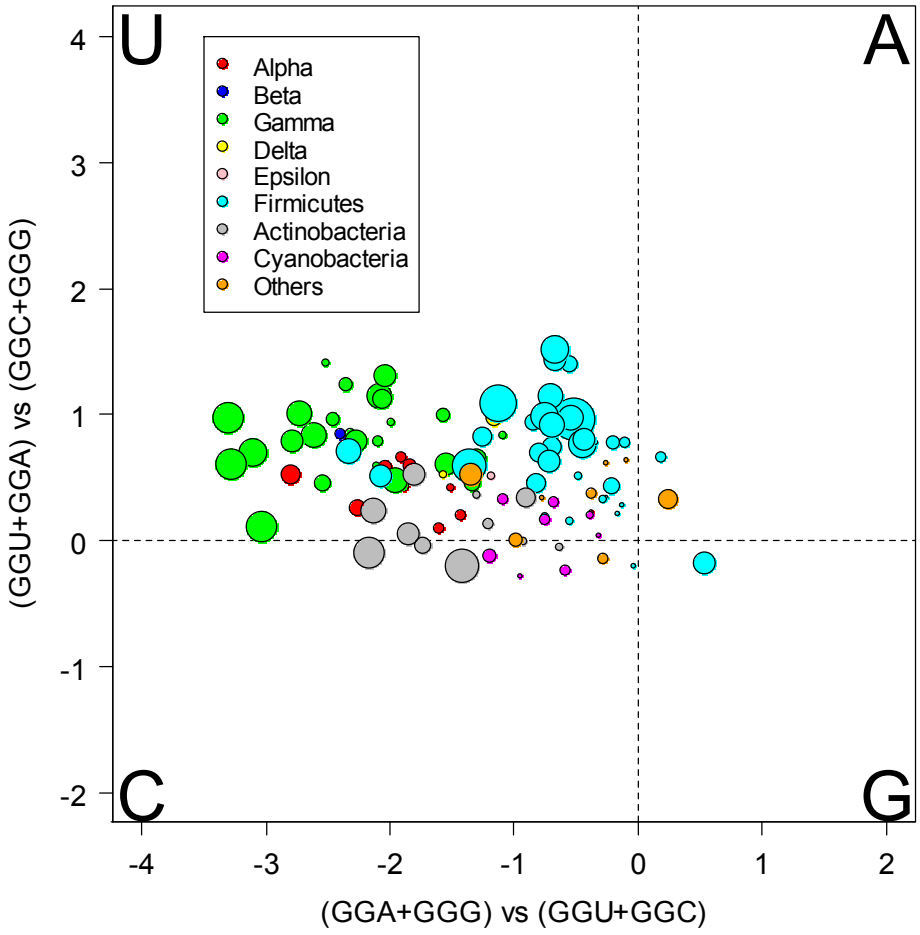

Figure 5-11 A plot showing the switching of selected codon usage bias for the amino acid glycine.

To consider changes in selected codon usage bias, codons used at significantly higher levels in the 40 highly expressed genes (*rplA-F*, *rplI-T*, *rpsB-T*, EF-*Ts*, EF-*Tu* and EF-*G)* as compared to the genome as a whole were deemed to be optimal codons. Significance was assigned after performing chi-squared tests and applying the Bonferroni correction using the same method as was used for the *B. bacteriovorus* genome in chapter 3. This analysis was done for all 160 bacterial genomes, although it must be stressed that optimal codons are only relevant if translational selection is the major factor influencing codon usage in the genome. In the previous chapter it was shown that 105 of the 160 genomes have an 'S' value indicating a significant degree of selected codon usage bias. It is possible however that the strength of selected codon usage bias statistic may miss genomes where selection is present due to the fact that it only takes into account four amino acids.

5.3.3.1 Alpha Proteobacteria

When the Alpha Proteobacteria (Figure 5-12) were examined it could be seen that they could be divided into two groups with respect to their phylogeny. The Rickettsiales showed a relatively low degree of selection and gave negative S_{WWY} values due to the frequent use of codons of the form WWU as opposed to WWC. The *Wolbachia*, *Ehrlichia* and *Anaplasma* genomes seemed to have codon usage heavily influenced by strong strand bias as NNG and/or NNU codons were the most frequently used. It was noticeable that the *Rickettsia* species appeared to have codon usage that was largely due to neutral mutations and so the codon usage of highly expressed genes would not be expected to differ from the genome as a whole. The fact that relatively few amino acids showed any kind of biased codon usage with respect to highly expressed genes reflected this. In the previous chapter some outline correspondence analysis results were presented that supports these observations with both *Ehrlichia species*, along with the *Anaplasma* species, showing that the primary axis in the correspondence analysis (CA) was associated with strand bias. The *Rickettsia conorii* genome also showed primary and secondary axes associated with GC3s and K3s respectively after a within-block correspondence analysis, used to eliminate

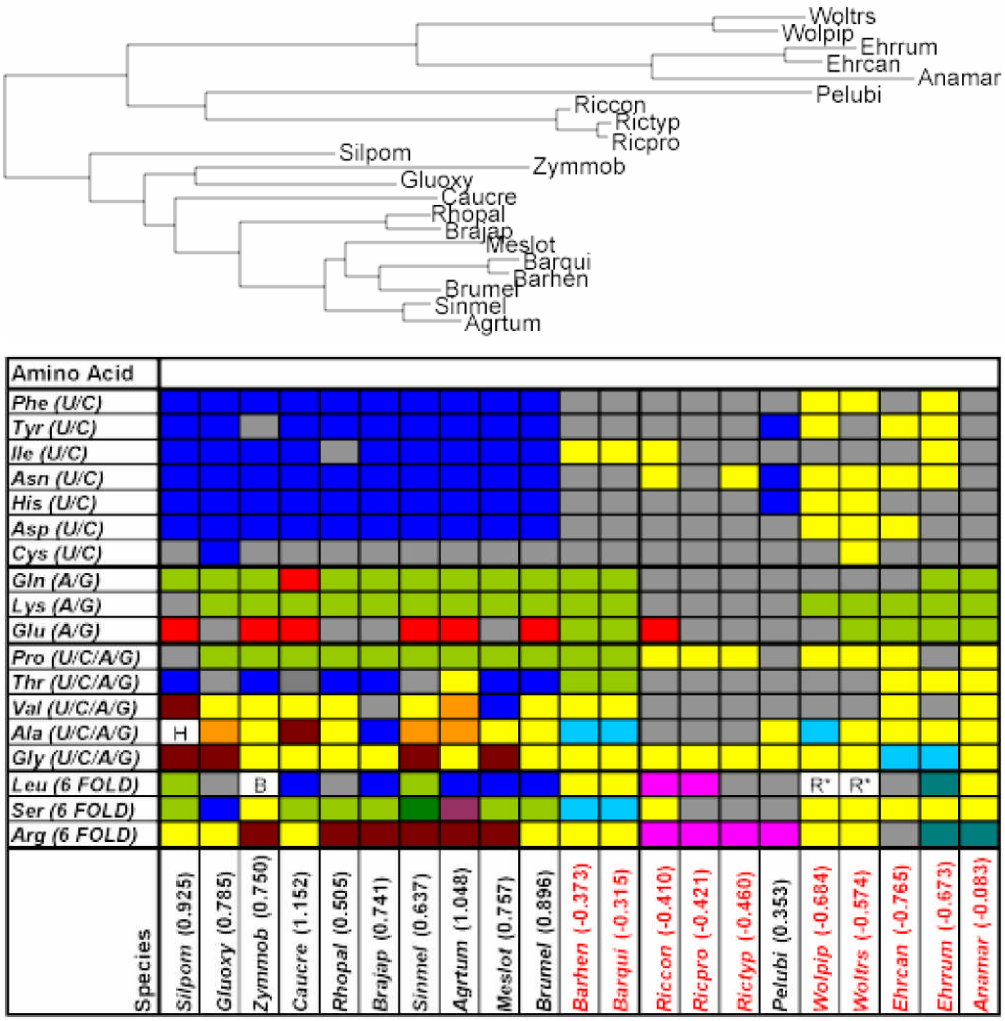

Figure 5-12 Figure illustrating preferred codon choice in the Alpha Proteobacteria.
Coloured squares indicate preference at the third codon position and standard single letter code is used to indicate codon usage where patterns are more complex. Reference to 'Other A', 'Other G' and 'Other C', refer to codon usage in sextets outside the quartet group. Similarly, a * indicates this when using single letter code. Genomes where S_{WWY} was not significant are coloured red.

the influence of rare arginine codons on the analysis. The genome of *Pelagibacter ubique* was the only one here to show any degree of selection, but the strength of selected codon usage bias was weak with only six codons showing possible evidence of selection.

Among the remaining Alpha Proteobacteria (i.e. other than the Rickettsiales clade) only two genomes were not assigned significant S_{WWY} values. These two Bartonella genomes appeared to show signs of strand bias with a strong 'preference' for codons of the form NNG and NNU. Such NNG/U preference showed up on these plots as in these genomes the highly expressed genes are almost all on the leading strand, which itself is G+U rich. The remaining genomes all appeared to show signs of translational selection with strong preference for the NNC codon in all codons of the form NNY apart from the rare amino acid cysteine, where no overall strong preference was identified. For amino acids with synonymous codons of the form NNR, the NNG codon is almost exclusively preferred over the NNA codon for the amino acids glutamine and lysine whilst for glutamate GAA is preferred. The four-fold degenerate amino acids proline and threonine show a clear preference for codons of the form NNG and NNC respectively, the only exception being *Agrobacterium tumifaciens* where NNU is preferred for threonine. Valine, alanine and glycine show mixed codon preference that, in some cases, adds the additional preference of NNC over the primary NNU preference. The trends observed here are supported by the switching plots but here one can see changes in selected codon usage bias for individual genomes more closely. This method also makes it possible to look at six fold degenerate amino acids. These amino acids also show preference for G and C ending amino acids with a mixture of NNG and NNC for leucine, a largely NNG preference for serine and some NNC preference over the predominant NNU preference for arginine.

5.3.3.2 Beta Proteobacteria

The main problem in identifying trends in optimal codon usage within the Beta Proteobacteria (Figure 5-13) was that selection is relatively weak with only one species showing a significant level of selection. The low amount of selection meant that little could be concluded from the switching plots but

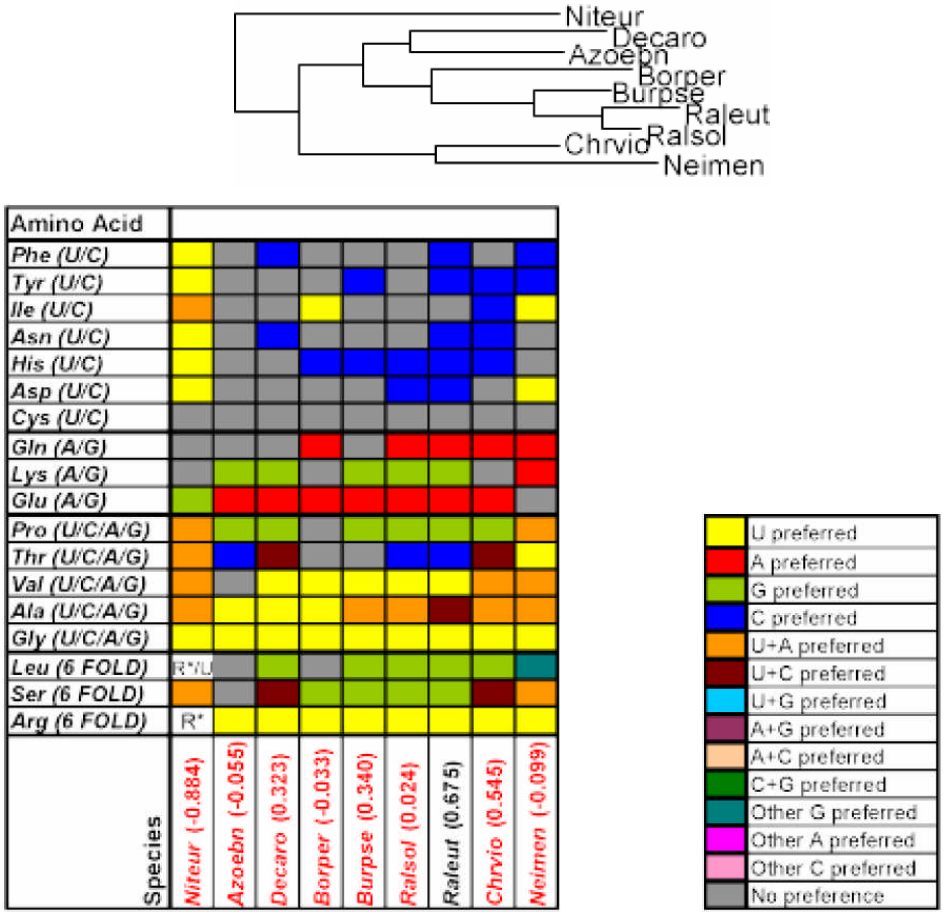

Figure 5-13 Figure illustrating preferred codon choice in the Beta Proteobacteria.

Coloured squares indicate preference at the third codon position and standard single letter code is used to indicate codon usage where patterns are more complex. Reference to 'Other A', 'Other G' and 'Other C', refer to codon usage in sextets outside the quartet group. Similarly, a * indicates this when using single letter code. Genomes where S_{WWY} was not significant are coloured red.

here it was possible to get a better insight into the codon usage of these genomes. The table shown here indicates that there may be some selection that has not been quantified by the S_{WWY} statistic. Although the NNY codons showed less selection than in other major clades the other codons showed some degree of selection. For the amino acids glutamine and glutamate CAA and GAA codons were predominantly the preferred codons. Upon closer inspection by correspondence analysis it seemed that, for the majority of these genomes, it was differences G+C content within the genome that was the major factor influencing codon usage bias with a correlation of at least 86% between GC3s and axis 1 for the Beta Proteobacteria in this study (correlations range from 0.87 in *Ralstonia eutropha* to 0.97 in *Chromobacterium violaceum*). This primary axis was seen to explain between 11% and 26% of the total variation and whilst some genomes analyses indicated that secondary axes may be influenced to some extent by selected codon usage bias it was difficult to separate these factors. This was compounded by the fact that often in these genomes the ribosomal genes were found in A+U-rich regions of the genome. It was therefore difficult to assess the exact interplay between selection and neutral forces in the Beta Proteobacteria. The plot for these species does not show any clear pattern of codon usage indicating that for many of these genomes several factors are influencing codon usage. The *Nitrososomonas europaea* genome is an example where the ribosomal genes are located in a particularly A+U rich area of the genome and in contrast to the high G+C content in the rest of the genome pick out A+U ending codons as being predominant, as discussed in Sharp *et al* 2005.

5.3.3.3 Gamma Proteobacteria

The Gamma Proteobacteria showed largely consistent codon preference (Figure 5-14) despite the large number of species in the clade. It appears from the table that the *Buchnera*, *Blochmannia* and *Wigglesworthia* species show no evidence of any kind of selected codon usage bias although some are influenced by strand bias, in the case of *Buchnera aphidicola* strain Bp and *Blochmannia floridandus*. The *Xylella* species also seemed to be influenced by strand bias, as has been previously reported (Sharp *et al.*,

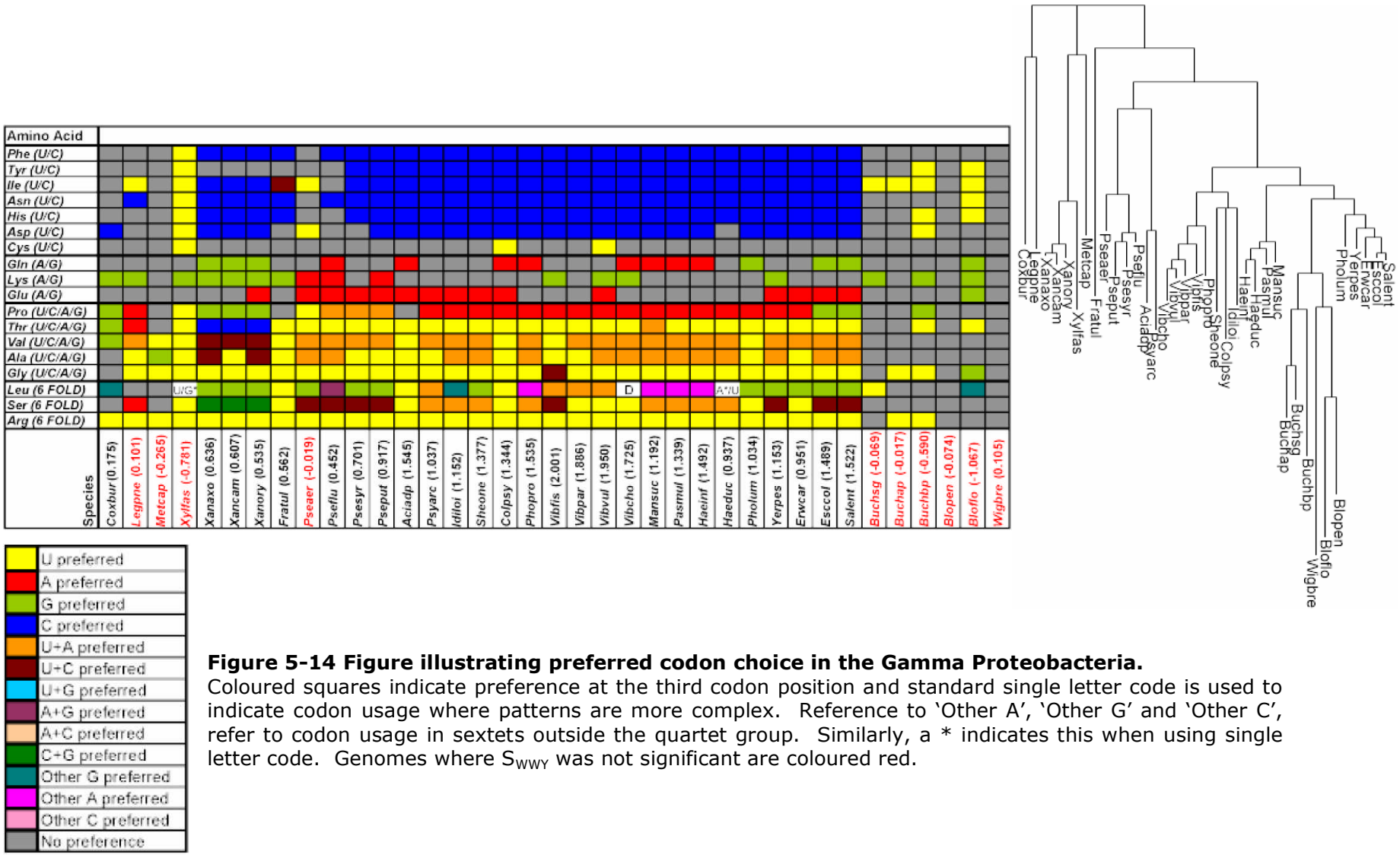

Figure 5-14 Figure illustrating preferred codon choice in the Gamma Proteobacteria.
Coloured squares indicate preference at the third codon position and standard single letter code is used to indicate codon usage where patterns are more complex. Reference to 'Other A', 'Other G' and 'Other C', refer to codon usage in sextets outside the quartet group. Similarly, a * indicates this when using single letter code. Genomes where S_{WWY} was not significant are coloured red.

2005). *Legionella pneumophila* also seemed to be influenced by variation in intragenomic G+C content, preferring NNA and NNU codons. In addition, no strong selection was evident in *Methylococcus capsulatus*. The *Pseudomonas aegrinosa* genome showed some evidence of selection, with codon usage similar to other *Pseudomonas* species, apart from in the two fold degenerate NNY codons. The *Pseudomonas* genomes are all G+C rich (GC3s: 71-87%) but the highly expressed genes in these species appear to be less G+C rich, indicating the possibility of selection. Previous codon usage analyses supported this view with evidence of horizontally transferred genes being the primary factor contributing to codon usage variation among genes but the secondary axis being due to translational selection (Grocock & Sharp, 2002). The clade as a whole did not seem to show switches in codon preference although strong selected codon usage bias was very apparent. The G+C rich *Xanthomonas* species show strong selected codon usage bias and also evidence for switches in codon usage with G+C ending codons being largely preferred as opposed to A+U ending codons for the rest of the Gamma Proteobacteria. The rest of the genomes show mainly A+U preferences or little preference at all. The *Vibrio* species use lysine AAG in three of the four species whilst the majority of other genomes show no preference for lysine. In addition the *E. coli* and *Salmonella enterica* genomes prefer NNG for glutamine and proline codons in contrast to the majority of Gamma Proteobacteria. Overall, for four-fold degenerate amino acids, NNU or NNU and NNA codons are preferred. The six-fold degenerate codons were also interesting with three clades of Gamma Proteobacteria (*Pseudomonas* species, a clade involving *E. coli*, along with the *Xanthomonas* clade) preferring CUG for leucine and some preferring UCC for serine.

5.3.3.4 Delta and Epsilon Proteobacteria

Due to the small number of sequenced genomes for the Delta and Epsilon Proteobacteria it was difficult to assess the extent of optimal codon switching in these clades (Figure 5-15). There were only two Delta and two Epsilon genomes that showed any sign of selection with little evidence of selection in the *Helicobacter* genomes along with the *Geobacter*

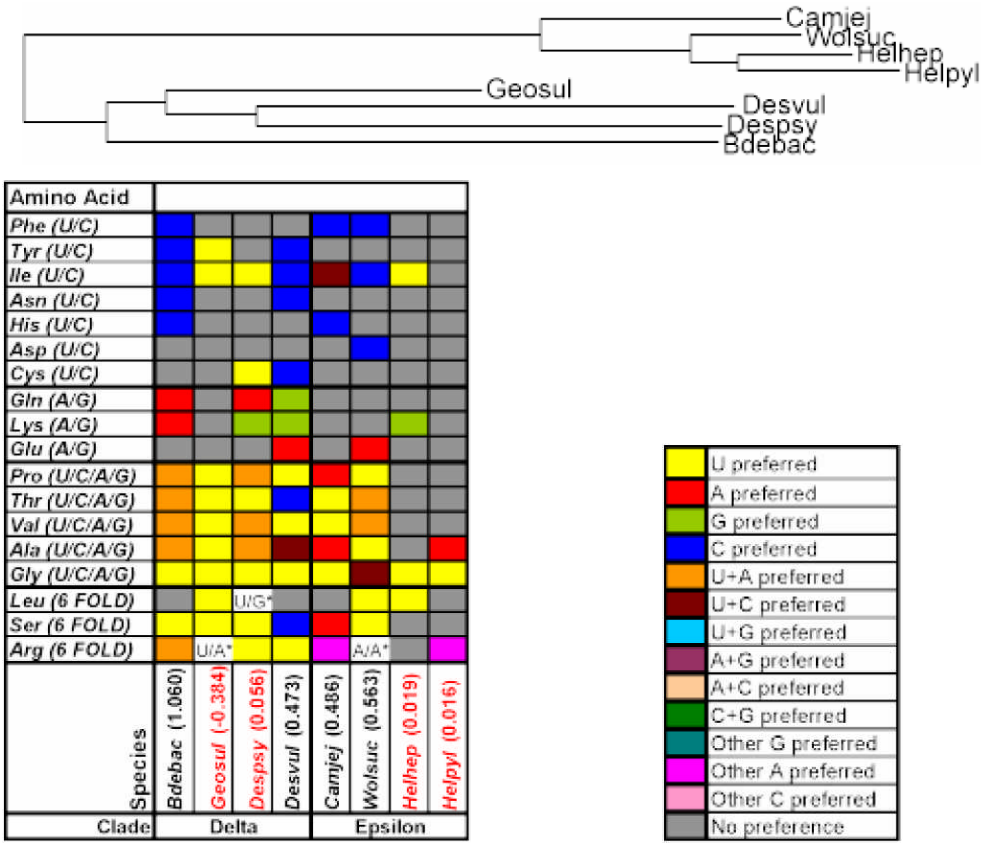

Figure 5-15 Figure illustrating preferred codon choice in the Delta and Epsilon Proteobacteria.
Coloured squares indicate preference at the third codon position and standard single letter code is used to indicate codon usage where patterns are more complex. Reference to 'Other A', 'Other G' and 'Other C', refer to codon usage in sextets outside the quartet group. Similarly, a * indicates this when using single letter code. Genomes where S_{WWY} was not significant are coloured red.

sulfurreducens and *Desulfotalea psychrophila* genomes. In chapter 3 of this thesis factors influencing the codon usage of *Bdellovibrio bacteriovorus* were discussed and it was shown that selection heavily influenced codon usage in this genome. One other Delta Proteobacterial genome, *Desulfovibrio vulgaris*, had an S_{WWY} value that was only just significant and the optimal codons identified did not match those of Bdellovibrio very closely. The Epsilon Proteobacterial genomes with significant S_{WWY} values, *Campylobacter jejuni* and *Wolinella succinogenes*, did not show strong signs of selection and again there were too few genomes to come to any reliable conclusions with regard to switches in optimal codon usage.

5.3.3.5 Firmicutes

The Firmicutes are a particularly well represented clade (Figure 5-16) and can be split into three major phylogenetic clades, *Bacilli*, *Clostridia* and Mollicutes. Overall, there was a high level of selection across the Firmicutes, with the most pronounced selection in the *Bacilli*. All *Bacilli* in the dataset were found to exhibit a significant degree of selection with S_{WWY} values ranging from 0.77 right up to 2.05. For the two-fold degenerate amino acids of the form NNY the NNC codon was almost exclusively optimal (NNC preference), apart from the rare amino acid cysteine, which showed relatively little codon preference. The amino acids glutamine, glutamate and lysine (with codons of the form NNR) showed a general preference for NNA with the only NNG preferred codons being the AAG codon for lysine in three L*actobacillus* species. Proline was seen to prefer CCA or a combination of CCA and CCU in most species whilst the other four-fold degenerate amino acids showed a major NNU preference with some species also using a NNA codon for the amino acids valine, alanine and threonine. The six-fold degenerate amino acid leucine shows a main NNU or NNA preference with several of the *Staphylococci* using the alternative UUA codon. Arginine showed a major NNU preference, with some species, most notably the *Bacillus* species using NNC additionally. In the case of serine NNU, NNA or a mixture of the two tended to be used with an NNA codon largely being used by the *Streptococcus* or *Lactobaccilus* species.

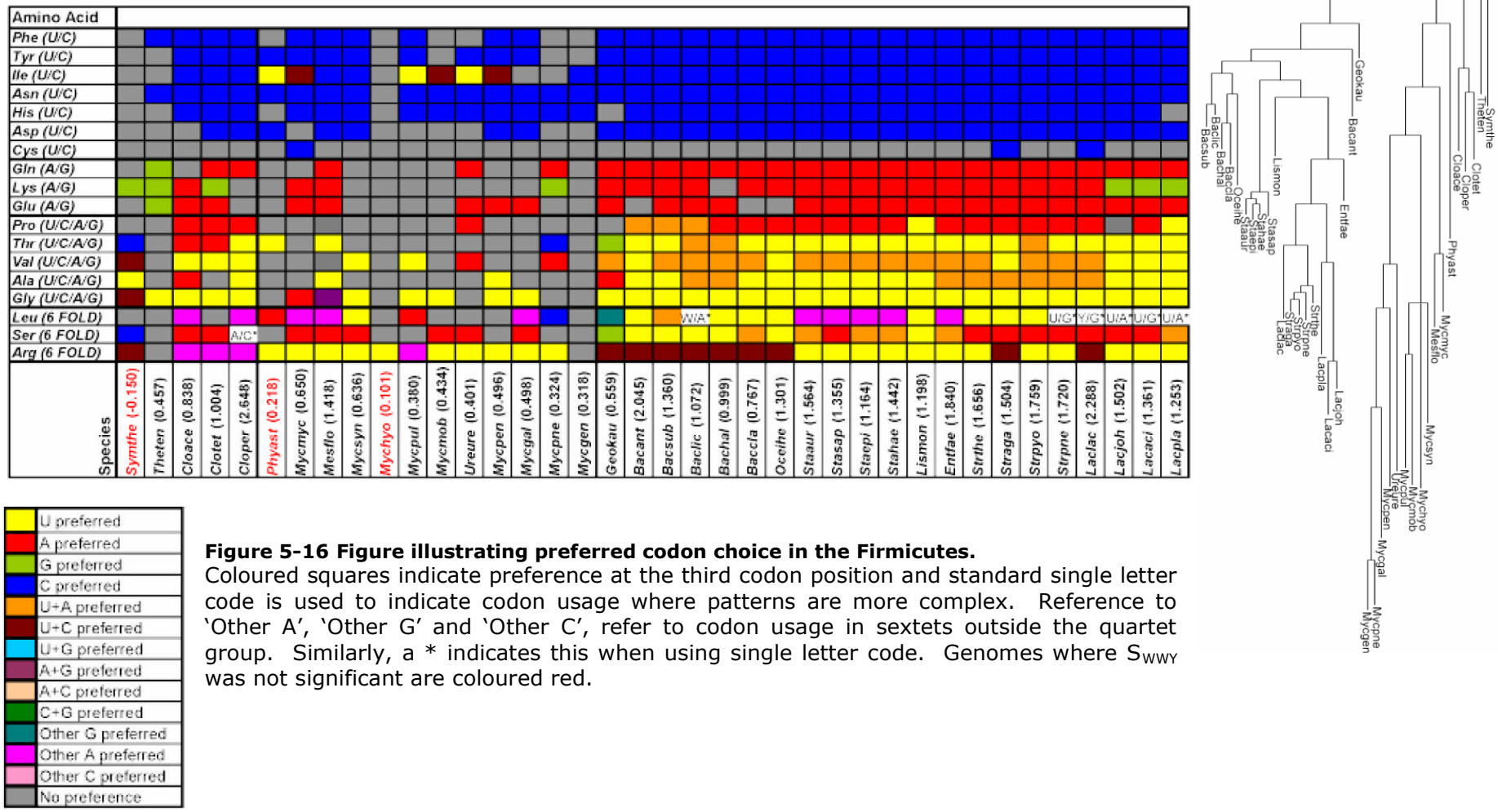

Figure 5-16 Figure illustrating preferred codon choice in the Firmicutes.
Coloured squares indicate preference at the third codon position and standard single letter code is used to indicate codon usage where patterns are more complex. Reference to 'Other A', 'Other G' and 'Other C', refer to codon usage in sextets outside the quartet group. Similarly, a * indicates this when using single letter code. Genomes where S_{WWY} was not significant are coloured red.

In the *Clostridia* there was evidence of selection with a strong preference for NNC in the NNY codons. The remaining two fold degenerates of the form NNR showed a general preference for NNA although *Clostridium tetani* preferred AAG for lysine and *Clostridium perfringens* showed no strong preference for lysine and glutamine. The four-fold degenerate amino acid, proline, showed a strong CCA preference in all three *Clostridiae* with CCU or CCA being preferred for the rest. As with the *Bacilli*, GGU always seems to be the preferred codon for glycine. The six-fold degenerate amino acids show a strong A preference with leucine and arginine preferring their alternative A-ending codons (UUA and AGA), possibly due to their A+U content.

The Mollicutes exhibited a much lower level of selection with a relatively sparse collection of optimal codons, although those that were preferred indicated a preference for U and A ending codons in the two fold degenerate amino acids with codons of the form NNR as well four-fold degenerate amin acids. The two-fold degenerates of the form NNY as usual showed a NNC preference except for some species where isoleucine used NNU as well. This feature was most unusual as one would not expect this codon to be used under selection as the NNC codon pairs better with the predominant tRNA for isoleucine GAU. Therefore given the uniformally low S_{WWY} values for these species (0.32-0.49) and the sparsity of significant optimal codons it is likely that selection is weak within the Mollicutes and other forces may be shaping codon usage here.

5.3.3.6 Actinobacteria

The Actinobacteria are a group of bacteria under strong selection with evidence of optimal codon switching also very apparent (Figure 5-17). Selection for NNC codons for the two-fold degenerate amino acids with codons of the form NNY was reasonably strong and the amino acids with codons of the form NNR all seemed to demonstrate a strong preference for NNG codons, in contrast to the Firmicutes in particular. The four-fold degenerate amino acid threonine showed a strong preference for the NNC codon in those Actinobacteria shown to be under selection. Proline seemed to show mixed preference from CCG in several Actinobacteria to CCU or

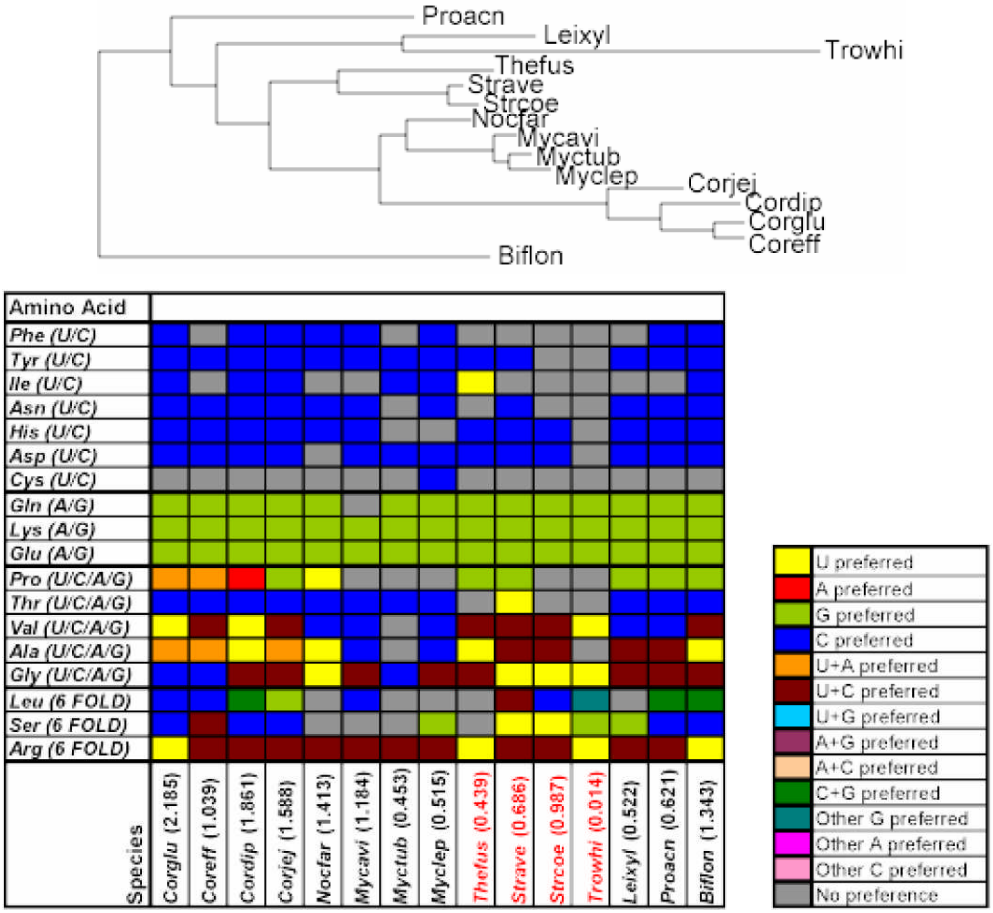

Figure 5-17 Figure illustrating preferred codon choice in the Actinobacteria.

Coloured squares indicate preference at the third codon position and standard single letter code is used to indicate codon usage where patterns are more complex. Reference to 'Other A', 'Other G' and 'Other C', refer to codon usage in sextets outside the quartet group. Similarly, a * indicates this when using single letter code. Genomes where S_{WWY} was not significant are coloured red.

CCA in others, most notably the *Corynebacteria*. A similar trend was also apparent for alanine where the *Corynebacteria* showed a GCU or GCA preference while other Actinobacteria preferred GCU or GCC. Valine and glycine showed preference for NNC and NNU mainly. The six-fold degenerate amino acids also showed a preference for C- and G-ending codons. Codon preferences here were in stark contrast to those of other major clades such as the Firmicutes and thus are evidence for a switch in optimal codon preference. Four genomes were shown to display no significant levels of selected codon usage bias. The S*treptomyces* genomes are extremely G+C rich and showed relatively high S_{WWY} values. Such S_{WWY} values for *S. coelicolor* and *S*. *avermitilis* are consistent with the expected value for a genome with number of rRNA operons and tRNA genes present in each genome (6 rRNA operons for both genomes and 63 tRNA genes in *S*. *coelicolor* with 68 tRNA genes in *S*. *avermitilis*). However the significance of these values could not be guaranteed as the broad range of S_{WWY} values for randomly selected genes in the genomes made patterns of codon usage variation difficult to interpret. Similar problems were evident with the *Thermobifida fusca* genome although the S_{WWY} value was lower in this case at 0.44. Codon usage in the *Tropheryma whipplei* is most probably influenced by strand bias given the strong NNG/U codon preference seen and a correspondence analysis on codon usage within this genome confirmed this with the primary axis in the correspondence analysis giving a correlation of 0.83 with K3s.

5.3.3.7 Cyanobacteria

There was limited resolution in this group of bacteria (Figure 5-17) but some trends could be identified. Firstly it was noticeable that *Synechocystis* PCC6803, both *Synechococcus* species and *Prochlorococcus marinus* strain MIT9313 all displayed some evidence of selection with the remaining genomes showing weak or no selected codon usage bias. Of these four genomes *Synechocystis* PCC6803 and *Synechococcus elongates* showed similar codon usage and *Synechococcus sp.* WH8102 and *Prochlorococcus marinus* strain MIT9313 showed similar codon usage, but the two groups differed. The latter group used a higher proportion of C-ending codons for the four and six fold degenerates and G-ending codons

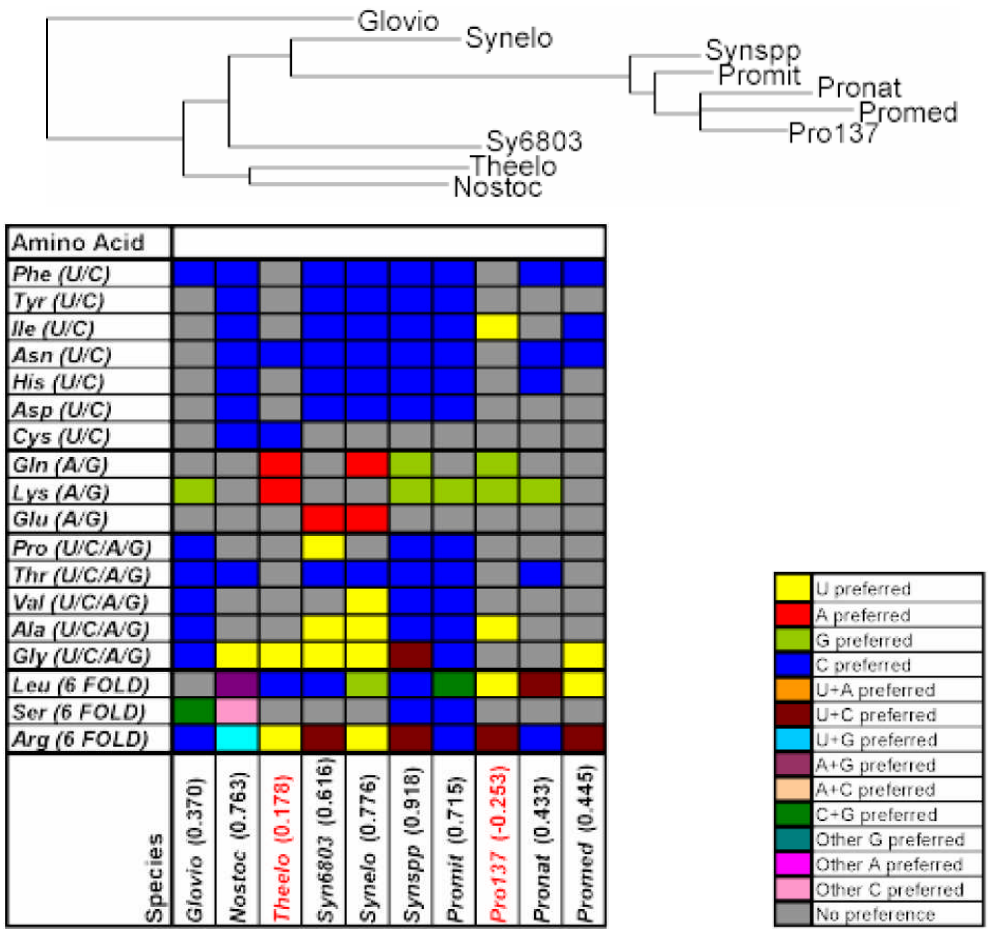

Figure 5-18 Figure illustrating preferred codon choice in the Cyanobacteria.

Coloured squares indicate preference at the third codon position and standard single letter code is used to indicate codon usage where patterns are more complex. Reference to 'Other A', 'Other G' and 'Other C', refer to codon usage in sextets outside the quartet group. Similarly, a * indicates this when using single letter code. Genomes where S_{WWY} was not significant are coloured red.

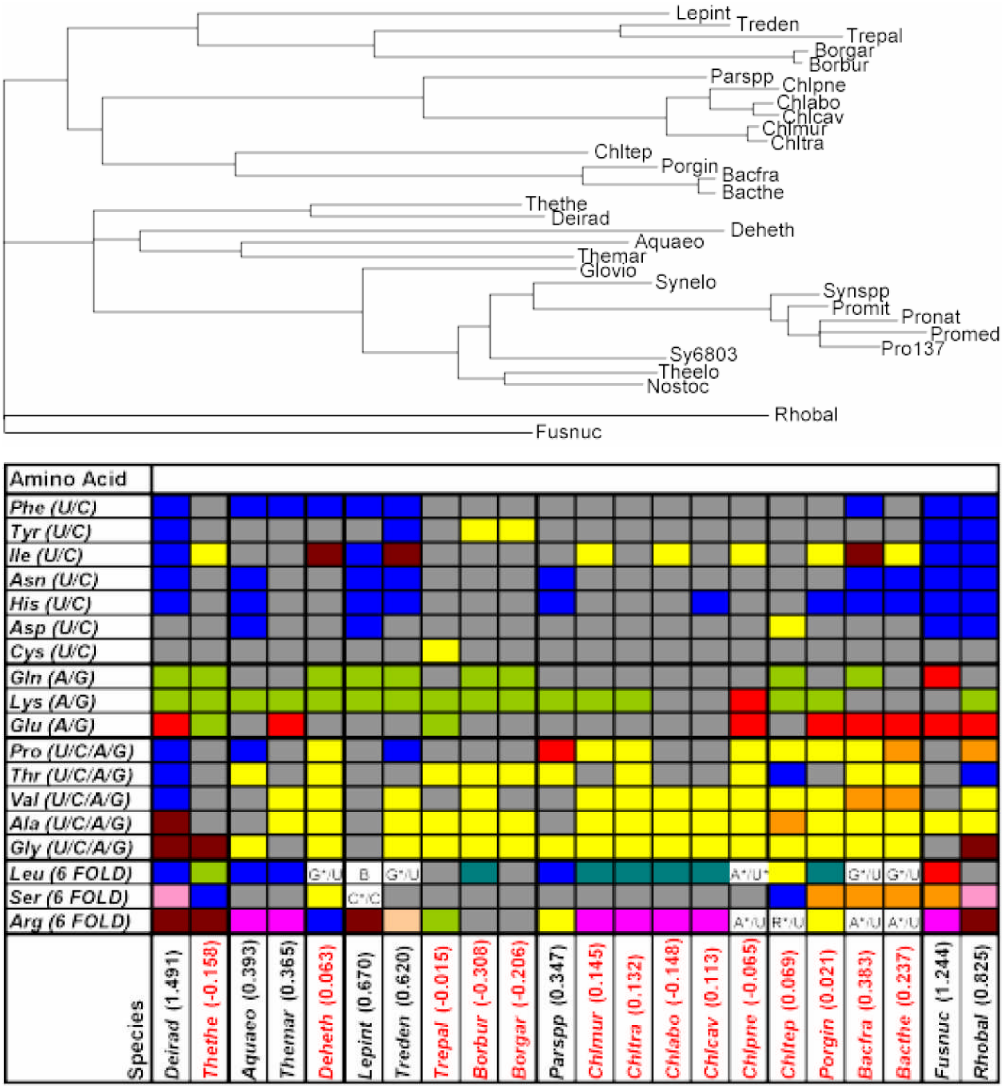

Figure 5-19 Figure illustrating preferred codon choice in the remaining bacterial genomes.
Coloured squares indicate preference at the third codon position and standard single letter code is used to indicate codon usage where patterns are more complex. Reference to 'Other A', 'Other G' and 'Other C', refer to codon usage in sextets outside the quartet group. Similarly, a * indicates this when using single letter code. Genomes where S_{WWY} was not significant are coloured red.

for two-fold degenerate amino acids such as lysine. In contrast the former group used more codons of the form NNU and appeared to prefer NNA codons for two fold degenerate amino acids with codons of the form NNR. With such few Cyanobacteria in this study these trends in codon usage bias are not highly reliable. It remains to be seen whether there is much strong selection with regard to codon usage in the Cyanobacteria which confuses the issue of switches in codon usage.

5.3.3.8 The remaining groups

The remaining species consisted of 22 additional genomes from various areas of the phylogeny. Only 8 of the 22 genomes showed a significant level of selected codon usage bias (Figure 5-19) and because of the wide distribution of these species general trends were hard to decipher. The most notable of these genomes however was *Deinococcus radiodurans* of the Deinococcus-Thermus phylum which displayed distinct codon usage patterns. This genome showed strong selected codon usage bias with an S_{WWY} of 1.491 and appeared to show a preference for NNC codons for four and six-fold degenerate amino acids and NNG codons for glutamine and lysine, with NNA preferred for glutamate. Relatively little can be inferred from these genomes due to their diverse nature and severe lack of species available. As more bacterial genomes are sequenced, for each of these individual clades, clearer patterns of codon usage should emerge.

5.3.4 Putting codon preference switching into an evolutionary perspective

The two previous methods of analysis of trends in selected codon usage bias found two major types of switch in codon usage bias. Firstly there were major switches in codon usage that encompassed large groups of bacteria. Secondly, there were smaller variations in selected codon usage bias that affected only one or two genomes and sometimes just one or two amino acids with the other amino acids maintaining the general trend. Strong evidence of optimal codon switching was collated from the previous analyses and was plotted onto the phylogeny (Figure 5-20) produced using the ribosomal protein sequences (*rplA-C, rpsB-C* and elongation factor *Tu*).

Evidence of large scale optimal codon switching can be seen for clades such as the Actinobacteria, the Alpha Proteobacterial clade that excludes the Rickettsiales, and possibly the Cyanobacteria, although the low number of Cyanobacterial genomes in the dataset exhibiting *strong* selected codon usage bias means that one cannot be confident in this case. Switches in codon usage were also observed for the G+C rich *Xanthomonas* species of the Gamma Proteobacteria. Small changes in codon usage could also be seen throughout the phylogeny and include switches for single amino acids such as glutamine and proline.

To try to understand how such switches in codon usage could take place various features of the genomes were examined to try search for any genomic features that could explain these switches in selected codon usage bias. The most striking correlation was that between optimal codon usage and genomic GC3s. Ancestral genomic GC3s at the base of each clade was estimated by a maximum likelihood approach using the package Continuous (Pagel, 1999). It can be seen in Figure 5-21 that red crosses, representing switches in codon usage bias, are primarily on branches leading to clades with high G+C contents. This indicates that this genomic feature may be an important factor influencing the switching codon preferences. This example is very striking for bacterial groups such as the Actinobacteria and the Alpha Proteobacterial clade where GC3s is particularly high and switches in selected codon usage bias have occurred. In the Gamma Proteobacteria the *Xanthomonas* species are also particularly G+C rich (estimated ancestral GC3s of 0.79) and again a switch in selected codon usage bias is evident.

The amino acids threonine, proline, lysine and glutamine appear to be the most likely to switch, with these amino acids having contrasting codon usage in the Actinobacteria, Alpha Proteobacteria and *Xanthomonas* species of the Gamma Proteobacteria (codons of the form NNG/C) as compared to the Firmicutes and remaining Gamma Proteobacteria (codons of the form NNA). The amino acid glutamate also switched in the Actinobacteria (to a GAG codon preference) although not in the Alpha Proteobacteria or *Xanthomonas* species. These switches in selected codon usage bias were

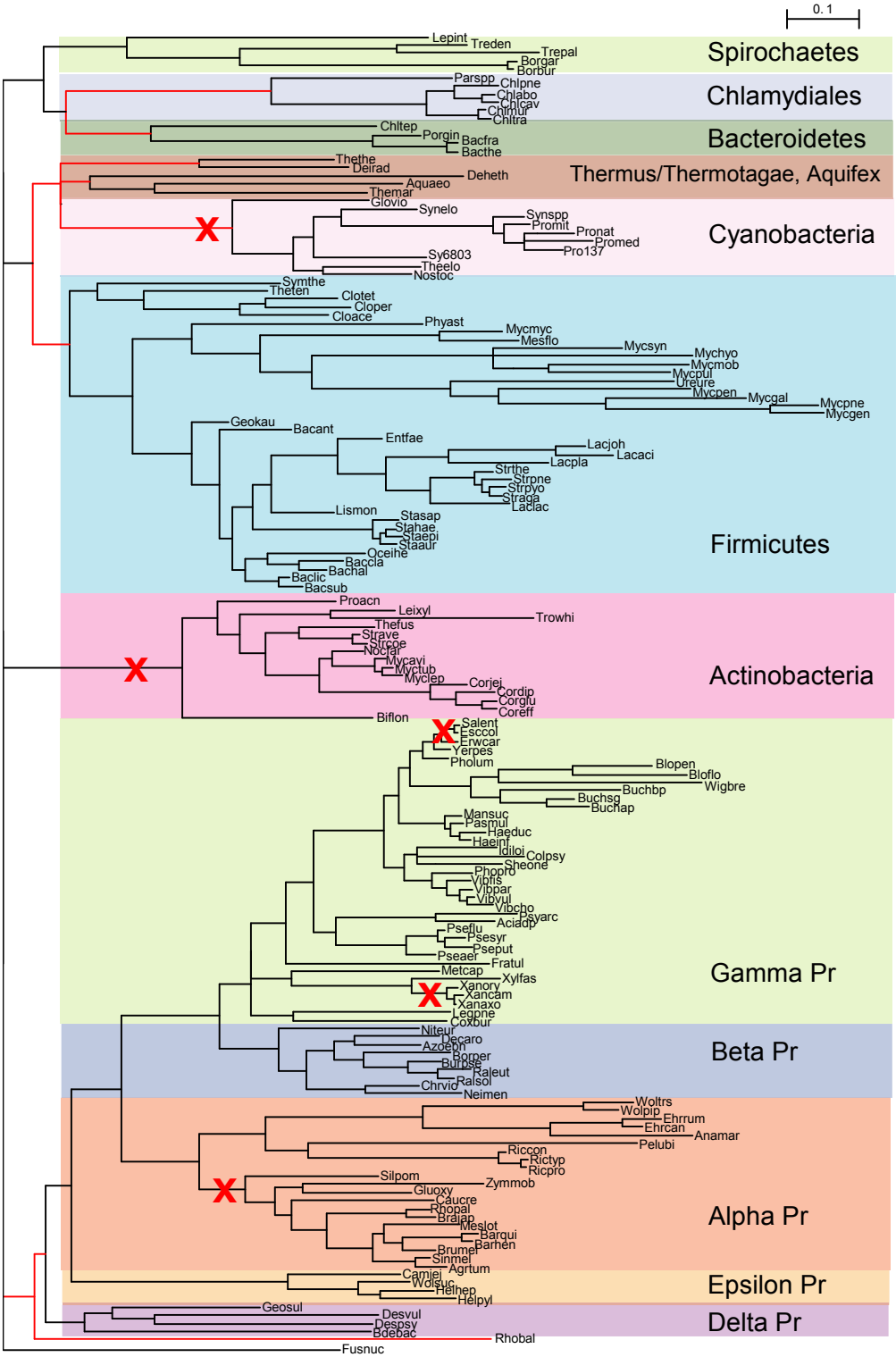

Figure 5-20 Complete phylogeny of the 160 genome dataset
Codon switches from A+U to G+C are marked as red crosses. Branches with less than 70% posterior probability are also marked in red whilst those with less than 60% posterior probability were collapsed.

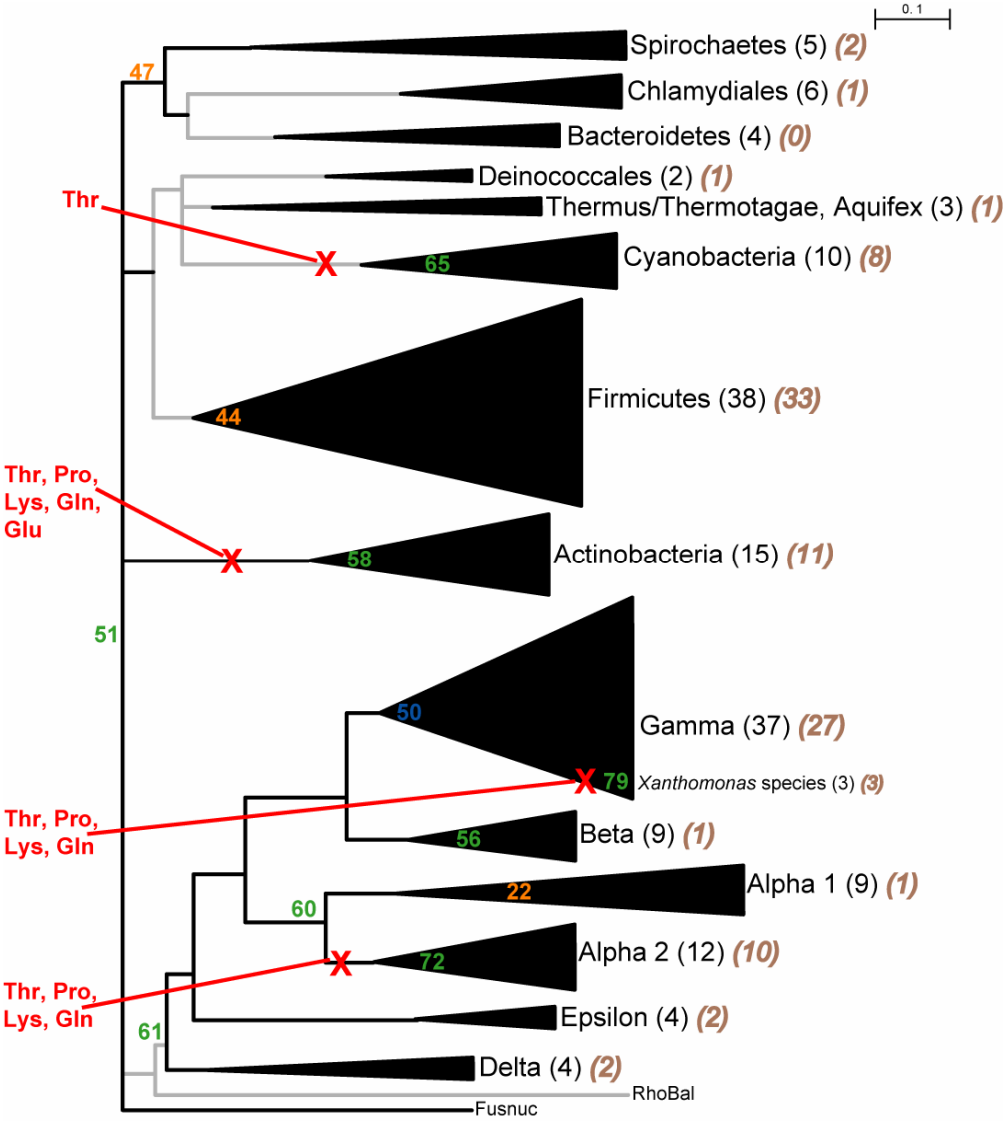

Figure 5-21 Outline phylogeny of the 160 genome dataset
Major switches in codon usage (for two and four-fold degenerate amino acids) are marked by red crosses and ancestral GC3s content shown for each major clade. The numbers of bacterial species in each clade are shown in brackets, the numbers in grey italics showing the number of species with significant levels of selection. Grey lines indicate branches with posterior probabilities less than 70%.

strongly correlated with changes in genomic G+C content. Of the six-fold degenerate amino acids the leucine and serine amino acids were possibly seen to switch codon preference (although not always to the same synonymous codon) whilst arginine sometimes additionally used a NNC codon under high genomic G+C content (in some Actinobacteria and Alpha Proteobacteria) along with the NNU codon.

However, smaller switches that affected the codon usage of only one or two amino acids were also seen to occur with no strong correlation to changes in genomic G+C content. Such switches were evident for the glutamine and proline codons in *E. coli* and *Salmonella enterica*.

5.4 Discussion

It can be seen from the analysis described above that selection is acting to influence codon usage across the bacterial phylogeny. However, the strength of this influence and the effect it has can be seen to vary. It appears that some changes in selected codon usage bias affect multiple amino acids whilst others affect just a few. This can be seen in Figure 5-21 where some amino acids such as threonine switch in all the cases shown whereas an amino acid such as glutamate only switches its codon usage at the base of the Actinobacterial clade. Additionally, some amino acids retain the use of one particular codon across all bacterial genomes under selection. Amino acids such as phenylalanine, tyrosine, isoleucine, asparagine, histidine and aspartate all have synonymous codons of the form NNY and in these cases the NNC codon appears to be almost universally preferred due to its exact base pairing with the GNN anticodon (the only anticodon available for NNY codons). Finally, many switches affect large groups of bacteria whilst others affect only a few closely related genomes.

Figure 5-21 is largely concerned with large scale changes in codon usage that appear to be correlated with mutational biases within a genome but other minor switches in codon usage have been discussed such as the switch in codon usage of proline in some of the Enterobacteria.

Here I look at what factors can be seen to be influencing codon usage patterns, to try to understand further how the phenomenon of switching codon preference can occur. The previous figure (Figure 5-21) showed how the large scale changes in selected codon usage bias were correlated with genomic GC3s and here I will discuss how such switches can occur.

5.4.1 The Shields hypothesis

As outlined in the introduction to this thesis, Shields suggested a model whereby a sufficiently strong switch in mutational bias could alter optimal codon preference without a relaxation of selection. He also suggested that genes under weak selection should change their codon usage in line with the change in mutational bias whilst genes under strong selection would maintain their optimal codon usage until the change in G+C content made the use of these optimal codons unsustainable. Such a situation would arise when the new G+C content of the genome would mean the old optimal codons were now too rare and as such disadvantageous.

Not only did Shields outline this model, he looked for evidence of such codon usage patterns in a variety of organisms. To see if the actual codon usages of highly expressed genes fit in with the pattern predicted, Shields plotted codon frequencies in highly expressed genes against mutational bias (estimated from codon frequencies in lowly expressed genes). Evidence of such patterns of codon usage did exist although the lack of sequence data meant that only 7 genomes were analysed. It is now possible to look to see if such patterns of codon usage are visible in the much larger 160 genome dataset.

5.4.2 Evidence for the Shields hypothesis

The predictions of the Shields hypothesis appear to be validated by the results shown in this chapter. It does indeed appear that changes in genomic G+C content over a critical level have caused switches in optimal codon usage in many major bacterial clades (Figure 5-21).

Another thing that was done with the data was to see whether the graphical models produced by Shields looking at the relationship between codon frequency and mutational bias fit with the data presented here. To test the predictions the codon frequencies of the amino acid glutamine, for the 160 genomes, were plotted against genomic GC3s as a measure of mutational bias. When the data was plotted, Figure 5-22, it appeared that the curve predicted by Shields was indeed supported by the data, the sigmoid curve has genomes under strong selection at both the top and bottom tails of the curve with presumably a quick switch in codon preference between the two states as brought about by a change in mutational bias. Species with lower S_{WWY} values should show a more gradual change in codon usage that is indicative of drift and this can also be seen on the top graph. Although 29 of the 35 genomes with S_{WWY} values above 1.25 appeared to show strong evidence of codon switching (frequency of CAG below 0.2 or above 0.8) six genomes did not fit as well with the prediction (Figure 5-22, bottom plot). *Salmonella enterica* and *Vibrio cholerae* were only slightly unusual (frequency of CAG at 0.79 and 0.22 respectively) but other species such as *Photobacterium profundum* could not be explained so easily. However, all six of the datapoints identified belong to the Gamma Proteobacteria and it was noticeable in the switching plot that there was no strong codon preference in the Gamma Proteobacteria for the amino acid glutamine. This was, therefore, the likely cause for the points located away from the two tails expected on the plot as whilst there was strong selected codon usage bias for many amino acids in these genomes this was less true for glutamine.

5.4.3 Codon switching on an adaptive landscape

The abrupt change of optimal codon usage would, of course, have a major impact on the genome. Not only would synonymously variable sites in highly expressed genes be affected but a shift in tRNA abundances would also be expected due to the high degree of co-adaptation between these two genomic features. Wright's "shifting balance" theory (Wright, 1977) considers how a movement across an adaptive landscape may occur from one co-adapted state to another. This movement may be initiated by many factors. One possibility identified by Wright was that a long-term drop in

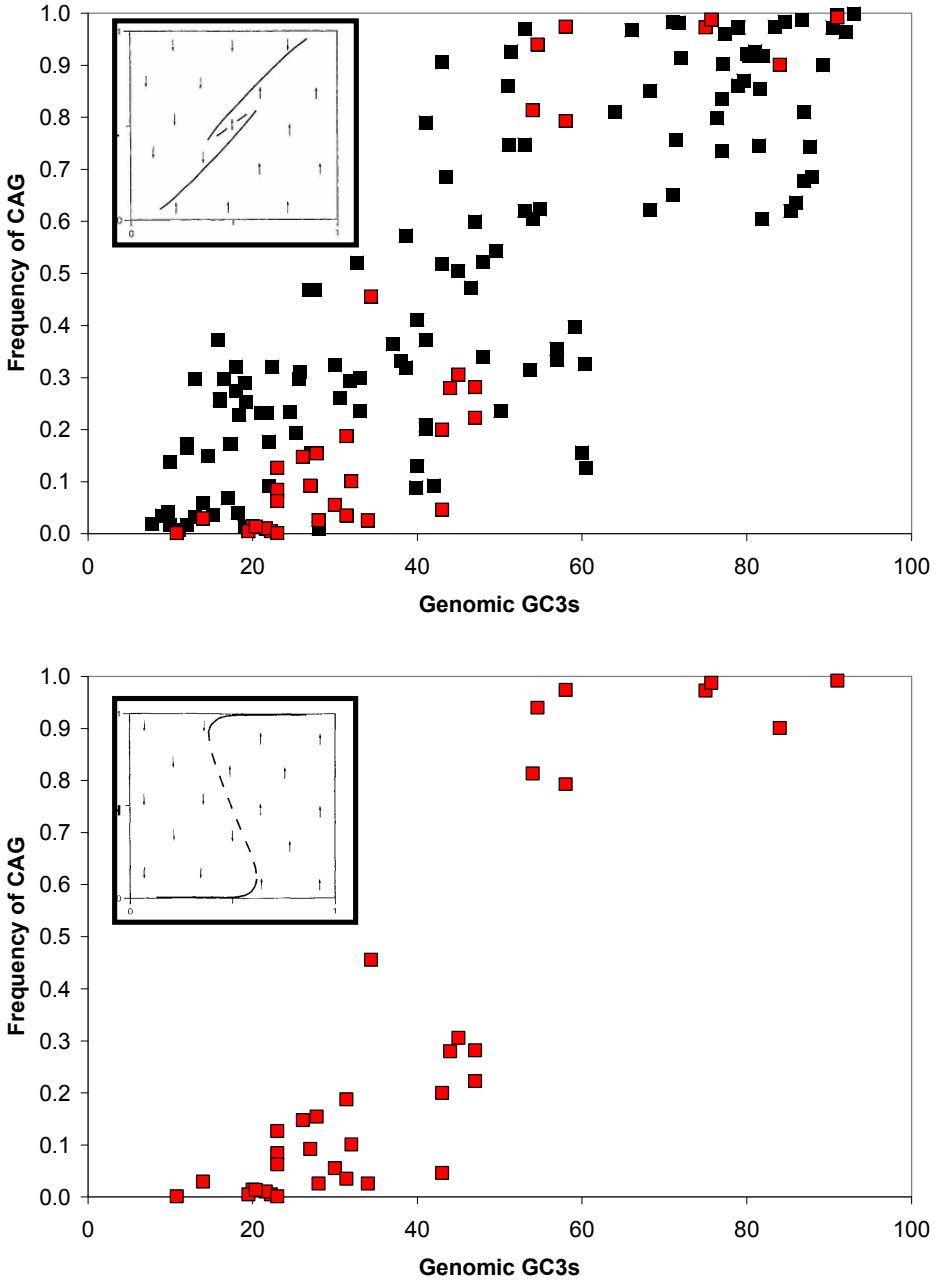

Figure 5-22 Codon switching curve for the amino acid glutamine. The frequency of the CAG codon in the 40 gene highly expressed dataset was calculated and plotted against the genomic GC3s value used to illustrate mutational bias within a genome. Genomes showing strong selected codon usage bias (SWWY value >1.25) are highlighted in red. Curves predicted by Shields are shown in the top left corner of each plot.

population size could initiate this movement by random drift. This could lead to randomization of codon usage away from their original codon preferences. A lowering of selective pressure for other reasons could also instigate this change and result in drift away from the original codon preferences. When selection was restored it could therefore be at a different point on the adaptive landscape but would have arrived there by neutral means. An alternative situation could occur if selection and population size remained constant then a change in mutational bias above a critical level could cause an abrupt change in codon usage preference without any stable intermediate steps.

It remains unclear whether reductions in population size occur regularly for unicellular organisms although this method may explain some switches in codon usage seen. With regard to the possibilities highlighted here, it is difficult to distinguish between the two methods from the data as they would both give effectively the same outcome. Therefore one can only speculate as to what may have happened but Shields' hypothesis, codon switching without a decrease in selection, does seem highly possible.

5.4.4 Explaining small scale changes in codon preference

The process described explains large scale changes in codon usage bias as a result of a switch in the mutation bias of a genome. However there are many examples, such as the use of the proline CCG codon in some Enterobacteria, where just a few amino acids are affected and are not correlated to genomic composition. It may be that such changes in selected codon usage bias are due to the acquisition of new tRNA genes allowing a new codon to be used. The process could, of course, occur in reverse with codon usage changes promoting tRNA acquisition but these changes in preference do not seem to be linked to genomic features such as G+C composition so it seems these switches are initially caused by neutral means but subsequently maintained by selection. In order to look further at small changes in codon preference such as these an examination of tRNA abundance is needed.

5.4.5 Why should only some amino acids show switches in optimal codon usage

Switches in codon usage do not occur across all amino acids at the same time. As has been reiterated through this chapter, amino acids with synonymous codons of the form NNY always prefer NNC as the anticodon of the tRNA used to decode these codons is exactly complementary, i.e. takes the form GNN. The ANN tRNA is very rarely seen due to the avoidance of A at the first tRNA position. This feature of tRNAs has been noted many times but no explanation for this is agreed upon. One possibility is that A avoidance is related to the stereochemical destabilization of the codon:anticodon duplex (Lim & Curran, 2001).

In contrast amino acids with synonymous codons of the form NNR seem to be able to switch in codon usage a lot more readily along with the four fold degenerate amino acids proline and threonine. These amino acids were where the bulk of the switching was evident in the dataset and can be seen in the Actinobacteria and Alpha Proteobacteria clades (where NNG or NNC codons were preferred). The other four-fold degenerates appeared to show a remaining preference for NNU codons, whilst the six fold degenerate amino acids showed ability to switch but with still a NNU underlying preference. Why these differences exist is unclear.

One possible difference could be the magnitude of selection for particular amino acids. The switching graphs showed that when switches take place the maximum S_{aa} values are often drastically different for specific amino acids. However, there does not appear to be a correlation between strength of S_{aa} and likelihood to switch. Additionally there does not appear to be a correlation with the frequency of use of an amino acid and its likelihood to be involved in optimal codon switching. In my opinion the potential key to why some amino acids are less likely to change their codon preference than others comes down to the accuracy and efficiency of translation. It is possible that some codons and tRNAs can cause problems either in the processivity of translation or may decrease the accuracy of

translation by introducing frameshifts or increasing the chance of an incorrect amino acid being incorporated into the growing polypeptide chain. It seems logical that this should be the main factor influencing codon preference as it seems that selected codon usage bias has the aim of increasing translational accuracy and efficiency. In order to look at this further it would be useful to look at the tRNA usage for each of these genomes to get a better idea of the overall picture with respect to tRNA abundance and codon usage.

5.4.6 Genomic composition bias in bacterial genomes

Genomic G+C content has been shown to vary widely among bacteria from 25% to 72% (Muto & Osawa, 1987; Ohama *et al.*, 1990; Ohkubo *et al.*, 1987). Much work has been concerned with the consequences of alterations in G+C content, as is the work presented in this chapter, however little is understood about how genomic base composition changes. It seems most likely that base composition evolves by neutral processes as suggested by Sueoka and Chen (Chen *et al.*, 2004; Sueoka, 1988). Arguments for the involvement of selection in the evolution of genomic G+C content have been put forward but there is little evidence for selective forces operating on a broad scale across the eubacteria. Some work has shown correlation between G+C levels and optimal growth temperatures in prokaryotes at the family level (Musto *et al.*, 2004) but the topic remains highly disputed and even this paper concedes that there is no correlation between genomic G+C content and optimal growth temperature at anything greater than the family level.

5.4.7 Optimal growth temperature and the bacterial common ancestor

Optimal growth temperature seems to be better correlated with ribosomal or transfer RNA G+C content and several studies have been done using this hypothesis to look at the nature of the bacterial ancestor and indeed the ancestor of all extant life forms (Galtier *et al.*, 1999). Results from these studies are often conflicting, with the main problems seeming to be a lack of resolution at the base of the bacterial phylogeny and the limited number

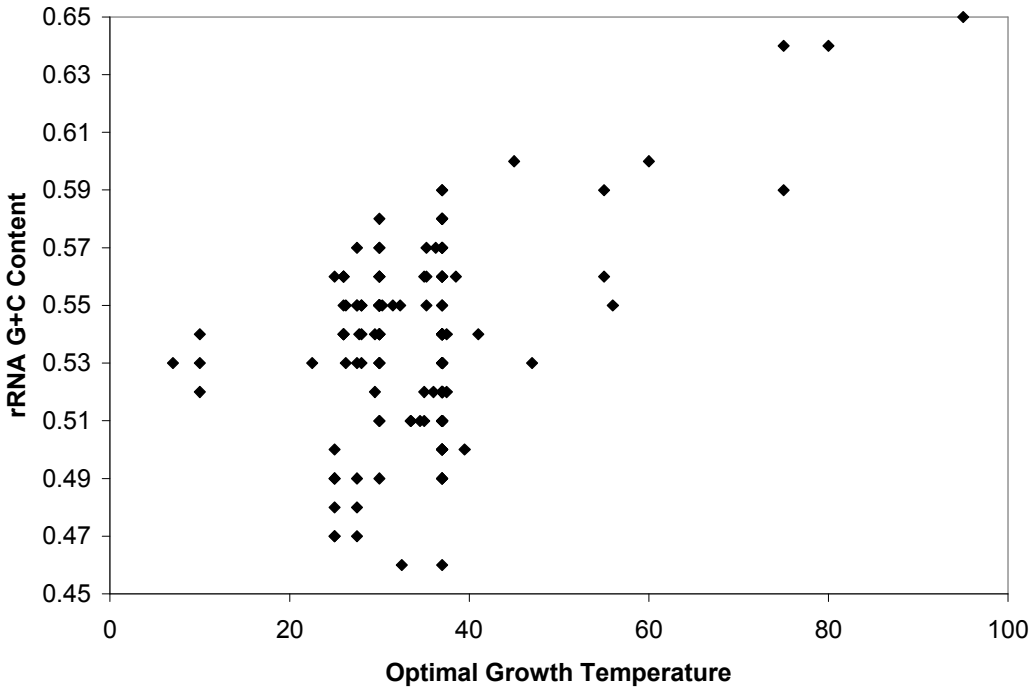

Figure 5-23 Plot of rRNA operon G+C content against optimal growth temperature.

of genomes used in the analysis. The dataset in this thesis shows some correlation between rRNA G+C content and optimal growth temperature (using optimal growth temperature data obtained from the recent Lobry paper (Lobry & Necsulea, 2006)) as can be seen in Figure 5-23. The phylogeny presented here suffers from the same issues in accuracy at its base and whilst the number of bacterial genomes included in my dataset is indeed large, relatively few thermophilic or hyperthermophilic genomes are present as few of these genomes have been fully sequenced.

In addition to the maximum likelihood estimations of clade GC3s content that were plotted onto my bacterial phylogeny I also calculated 16S rRNA G+C contents and estimated values for the base of the major clades. These values are not included here; however, as G+C values were not extreme or wide ranging as they only ranged from 50% to 56% G+C content. The main problem again was the lack of thermophilic and hyperthermophilic bacteria in my dataset. It therefore seems that the data here is insufficient to support either argument.

5.5 Conclusions

In this chapter it has been demonstrated that optimal codon switching is an important factor in the evolution of bacterial genomes. In particular it has been shown that patterns of selected codon usage bias switch to the use of NNG/C codons for major bacterial groups such as the Alpha Proteobacteria and the Actinobacteria, and that these switches in codon usage appear to be correlated with changes in mutational biases affecting genomic G+C content. In the next chapter I intend to discuss the relationship between patterns of selected codon usage bias and tRNA abundance to try and further understand the effects of directional selection upon these two highly co-adapted genomic features.

Chapter 6:
The relationship between optimal codon switching and highly co-adapted tRNA species

6.1 Introduction

The previous two chapters were largely concerned with factors affecting both the strength and direction of selection with regard to codon usage in bacterial genomes. Given the highly co-adapted nature of selected codon usage bias and tRNA abundance it made sense to further the study to see whether switches in codon usage bias that were found in the previous chapter were reflected with regard to tRNA abundances. The abundance of tRNA in the cell has been shown to be highly correlated with tRNA gene copy number (Ikemura, 1985; Kanaya *et al.*, 1999) and this approximation will be used throughout this chapter.

6.1.1 Models to predict the association tRNA abundances and selected codon usage

Previous work looking at codon:anticodon preferences summarised three models to explain codon and tRNA choice (Rocha, 2004).

6.1.1.1 Perfect match model

The perfect match model predicts that the optimal codon should be the one that makes a perfect codon:anticodon interaction with the most abundant anticodon species for a particular amino acid. The perfect match model assumes Watson-Crick pairing between codon and anticodon without modification to the tRNA species. Such an interaction is thought to increase the accuracy/efficiency of translation and was seen by Ikemura in his work on patterns of codon usage and tRNA abundance in *E. coli* (1981), where he saw that 'optimal' codons were used at higher frequencies in highly expressed genes and correlated to the most abundant tRNAs in the cell.

6.1.1.2 Frequency model

The frequency model predicts that the optimal codon should be the one that can be decoded by the largest number of tRNA species. The optimal codon used matches the model if it corresponds to the codon that maximizes the number of tRNA species with which it can interact (as outlined in Rocha, 2007).

6.1.1.3 Stability model

The stability model predicts that the optimal codon should be one that avoids either very strong or very weak codon:anticodon interaction. The model predicts that in a codon where the first two bases are 'strong' (G or C) the final base should be 'weak' (A or U). Similarly the reverse is true so that optimal codons should take the form WWS or SSW but never WWW or SSS. Where the first two bases in a codon take the form WS or SW no prediction is made. Such a model was suggested by Grosjean and Fiers (1982).

6.1.2 The modification of tRNAs

Posttranscriptional modifications to tRNA genes can affect the specificity of the tRNA species (Table 6-1). It is not possible to tell whether a tRNA has been modified by examining genome information alone. Therefore, it is difficult to be sure whether tRNA species identified in the dataset are unmodified and pair with their complementary codons (see table for possible modifications to normal codon:anticodon pairing rules).

6.2 Materials and Methods

6.2.1 Obtaining tRNA abundance data

For the sake of consistency tRNA gene copy number data was obtained from the tRNAscan-SE Genomic tRNA Database (http://lowelab.ucsc.edu/GtRNAdb/). This database was created using the tRNAscan-SE software package on completely sequenced bacterial genomes and is maintained by Todd Lowe. The tRNAscan-SE program (Lowe & Eddy,

Crick's wobble rules		Modified rules	
Anticodon	**Codon**	**Anticodon**	**Codon**
G	U, C	G	U, C
C	G	C	G
		K^2C	A
A	U	A	U, C, G > A
U	A, G	U	U, A, G > C
		xm^5s^2U,	A > G
		xm^5Um,	
		Um, xm^5U	
		xo^5U	U, A, G
		I	U, C, A

Table 6-1 Table illustrating pairing rules for third codon position
As proposed by Crick (Crick, 1966) and expanded to include modified bases. k^2C, lysidine; xm^5s^2U, 5-methyl-2-thiouridine derivatives; xm^5Um, 5-methyl-2'-O-methyluridine derivatives; Um, 2'-O-methyluridine derivatives; xm^5U, 5-methyluridine derivatives; xo^5U, 5-hydroxyuridine derivatives. Taken from Cochella & Green, 2004

1997) combines several tRNA finding algorithms with the main one for our purposes being tRNAscan 1.3. Candidate genes identified using these methods are then passed to the RNA covariance analysis package, Cove, for further analysis to remove false positives.

6.2.2 Deriving consensus tRNA complements across bacterial clades

In order to make sense of the tRNA abundances data extracted from the tRNAscan-SE Genomic tRNA Database "consensus" tRNA complements were derived for each of the major clades. This was done by comparing the tRNA abundances in each of the bacterial species of each clade to create a tRNA abundance that is typical of that clade. Typically, this approximated to the average number of each tRNA species among the bacteria in a particular clade.

6.2.3 Within-block correspondence analysis

Within-block correspondence analysis (Benzécri, 1983; Lobry & Chessel, 2003) was performed on the raw tRNA abundance data (i.e. gene copy number) for the 160 bacterial genomes used previously in this thesis (Appendix D) extracted from the tRNAscan-SE Genomic tRNA Database. This method was performed in a similar way to the within-block analyses carried out on codon usage data from the *Bdellovibrio bacteriovorus* genome in chapter three. The *R* statistical environment (http://www.r-project.org/) was used and in particular the *ade4* package (Thioulouse *et al.*, 1997) just as it was in the *Bdellovibrio* analysis. However, instead of synonymous codons being grouped together, isoaccepting tRNAs were grouped together in this analysis.

The tRNA species for the amino acid isoleucine were excluded from the correspondence analysis due to problems with the annotation of such tRNAs in the database. In particular, many isoleucine tRNA species with the anticodon CAU are missannotated as methionine tRNAs. However, this is not a problem for the analysis since the optimal codon for isoleucine

appears to be invariant across bacteria; with AUC always the optimal codon.

6.3 Results

6.3.1 Correspondence Analysis

The within-block correspondence analysis was performed to enable sense to be made of the huge amount of data extracted from the tRNA database. The technique was employed as an initial investigation to pick out the major trends influencing tRNA abundances.

6.3.1.1 Axis 1 correlates with G+C content

The primary axis found through the data explained 33.5% of the variation within the dataset. It can be seen that G+C rich genomes, such as those of the Actinobacteria and the Alpha Proteobacteria with significant selected codon usage bias, are all clustered to the right hand side of axis one (Figure 6-1). In contrast A+U rich genomes such as the Firmicutes cluster to the left hand side of the plot. Genomes such as the G+C rich *Xanthomonas* species of the Gamma Proteobacteria also cluster to the right hand side, consistent with their G+C content. The distribution of species across the axis appeared to result from changes in the relative abundance of different tRNAs in response to genomic G+C content, with a correlation of 0.75 between axis one and genomic GC3s. Therefore it appeared that the switches in selected codon usage bias as a result of G+C content are likely to be correlated with changes in tRNA abundances. An examination of the anticodons responsible for axis one made it clear that the primary anticodon species responsible for the pull to the right on axis one were CNN anticodons. The anticodons for the opposing pull were primarily UNN anticodons, although the threonine anticodon AGU and the leucine anticodon AAG were at the extreme left (Figure 6-2).

6.3.1.2 Axis 2 is influenced by arginine codons

The source of variation on axis two is primarily from the use of two tRNA species, with anticodons GCG and UCG, which are complementary to the arginine codons CGC and CGA. The genomes picked out by this axis appear

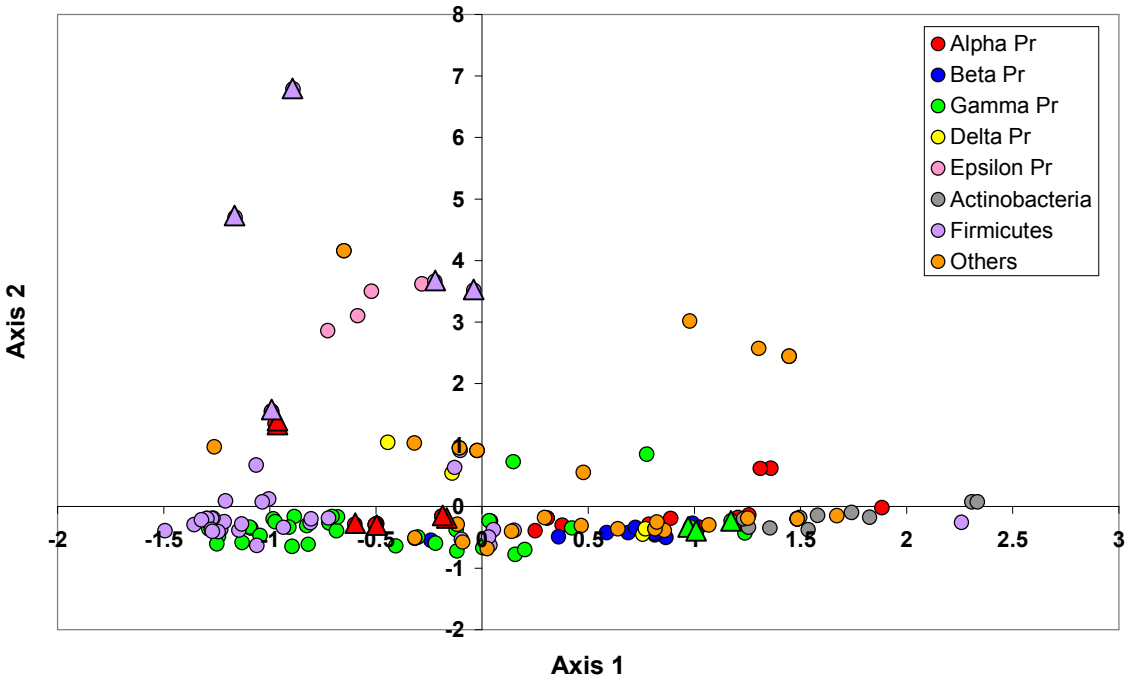

Figure 6-1 Within-block correspondence analysis of tRNA abundance data. Bacteria from different phylogenetic groups are colour-coded (see inset key). Points plotted as triangles are the Rickettsiales clade of the Alpha Proteobacteria, Xanthomonales clade of the Gamma Proteobacteria and the Mollicutes clade of the Firmicutes.

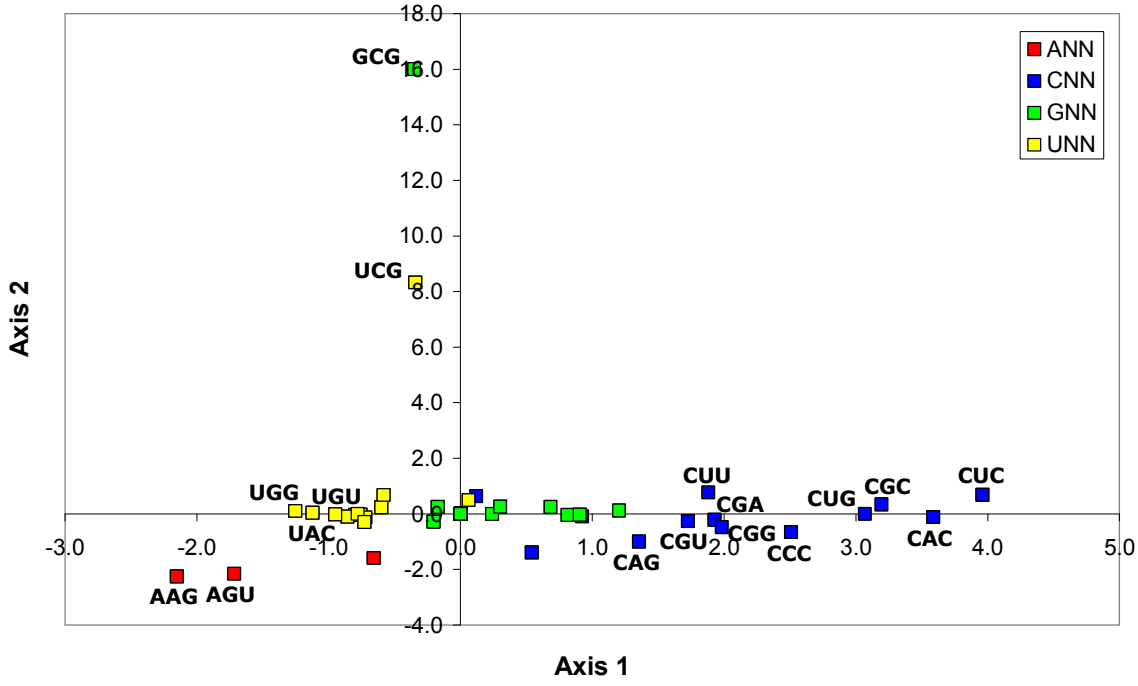

Figure 6-2 Anticdodon plot showing anticodons responsible for the trends in the within-block correspondence analysis of tRNA abundance data.

to be primarily from the Epsilon Proteobacteria, the Spirochaetes *Treponema* (*Treponema denticola* and *Treponema pallidum*) and *Borrelia* (*Borrelia burgdorferi* and *Borrelia garinii*), and the *Mycoplasma* species of the Firmicutes. Many of these genomes show no or relatively low selected codon usage bias and those exhibiting selection do not show optimal codons corresponding to these tRNAs. It appears that this axis is therefore strongly influenced by these rare tRNAs (GCG anticodon seen in just 9% of bacterial species in the dataset and UCG in 21%) and is not related to the primary focus of this chapter regarding codon usage switching and corresponding changes in tRNA abundance.

Similarly, axes three and four also appeared to be influenced by just a few rare tRNAs. The tertiary and quaternary axes accounted for 10.5% and 6.1% of the total variation within the dataset respectively. However, these axes were not deemed to be relevant within the context of this thesis due to the primary cause being the presence of rare tRNA species.

6.3.2 Codon switching patterns and the corresponding tRNA complements

In the previous chapter switches in optimal codon usage were seen at the bases of major bacterial clades (Figures 5-20 & 5-21). In particular, between the Alpha Proteobacteria exhibiting selected codon usage bias, the *Xanthomonas* species of the Gamma Proteobacteria, the Actinobacteria and *Salmonella enterica* and *E. coli* (and to some extent the Cyanobacteria) on one hand and the Firmicutes and majority of the Gamma Proteobacteria on the other hand. These switches in codon usage were seen to correlate with genomic G+C content. The within-block correspondence analysis of the tRNA abundance data showed that tRNA usage also correlated with genomic G+C content and so it is likely that codon usage switches and tRNA abundance changes may be correlated with each other and brought about by a change in genomic G+C content. In this section the tRNA abundances corresponding to major switches in codon usage were examined. In order to do this the complement of tRNA species present, for each genome with significant selected codon usage bias, in each of the major bacterial clades

was examined. Next, 'consensus' tRNA abundances representative of each clade were produced to give an indication of the ancestral state (Figure 6-3 and 6-4). Full tRNA abundance data can be found in appendix D.

The seven two-fold degenerate amino acids with codons of the form NNY were seen in the previous chapter to prefer the NNC codon over the NNU alternative across all bacteria with no switching of codon preference observed. The tRNA abundance data showed that the only tRNA species present to decode these amino acids have anticodons of the form GNN, thus explaining the universal preference for NNC codons (Figure 6-3). The same is indeed true for all bacterial species examined in this thesis and so these NNY codons will not be discussed again.

Similarly, consensus tRNA gene complements were largely consistent across all the bacterial species examined for the three six-fold degenerate amino acids leucine serine and arginine (Figure 6-4) and little evidence of codon switching was seen in the previous chapter (Figure 5-21). For the amino acid leucine all possible isoaccepting tRNA species were present except for tRNA species with AAG anticodons, which were never used and the tRNA species with anticodon of the form CAG that was not present in the Firmicutes only. The amino acid serine showed similar patterns with regard to its consensus tRNA gene complements with anticodons of the form ANN not present and all other tRNA species used (GCU, GGA, UGA and CGA) except in the case of the Firmicutes and Gamma Proteobacteria where tRNA species with CGA anticodons were not present. The amino acid arginine was the only amino acid where an anticodon of the form ANN was seen, and indeed often the most abundant isoaccepting tRNA species for arginine. The same four anticodons were seen in the consensus tRNA gene complements for all clades for arginine and these were UCU, CCU, ACG and CCG.

6.3.2.1 Gamma Proteobacteria (main clade)

In the case of two-fold degenerate amino acids with synonymous codons of the form NNR (Gln, Lys and Glu) the main Gamma Proteobacterial clade

(excluding the Xanthomonales as well as *E. coli* and *S. enterica* discussed separately) were seen to largely prefer NNA codons over the NNG alternative. When looking at the tRNA consensus for this clade, the only tRNA species seen in these Gamma Proteobacteria had an anticodon of the form UNN which was complementary to the NNA codon preference observed. As well as no tRNA species with anticodons of the form CNN being present it was notable that the tRNA species with UNN anticodons that were present were in multiple gene copy number (Figure 6-3). This may be indicative of optimization of the translational machinery using the perfect match model.

The four-fold degenerate amino acids were seen to prefer CCA codons for proline and codons of the form NNU (with some NNA preference) for the other amino acids. The consensus tRNA abundance for the main Gamma Proteobacterial clade showed the presence of tRNA species with anticodons of the form GNN and UNN. The tRNA species with UNN anticodons were seen to be often in multiple copies whilst the tRNA species with GNN anticodons were usually seen in single copy number (except for the amino acid glutamine where the situation was reversed).

6.3.2.2 Firmicutes

The A+U rich Firmicute clade was seen, in the previous chapter, to prefer NNA codons over NNG for the amino acids glutamine, lysine and glutamate. Similarly the consensus tRNA abundances for this clade support such codon usage with complementary tRNA species with anticodons of the form UNN preferred. These tRNA species were also seen to be present in multiple copies, thus increasing the abundance of these tRNAs in the cell (again supporting the perfect match model).

When the four-fold degenerate amino acids were considered the Firmicutes were seen to prefer codons of the form NNA or both NNU and NNA. Transfer RNA gene complements showed that the primary tRNA species seen had anticodons of the form UNN and these were found in multiple

Clade	Amino Acid						
	Phe	Tyr	Ile	Asn	His	Asp	Cys
Alpha Pr	1 GAA	1 GUA	2 GAU	1 GUU	1 GUG	2 GUC	1 GCA
Gamma Pr	1 GAA	1 GUA	3 GAU	2 GUU	1 GUG	2 GUC	1 GCA
Xanthomonales (Gamma Pr)	1 GAA	1 GUA	2 GAU	1 GUU	1 GUG	2 GUC	1 GCA
Escol, Salent (Gamma Pr)	2 GAA	3 GUA	3 GAU	4 GUU	1 GUG	3 GUC	1 GCA
Actinobacteria	1 GAA	1 GUA	1 GAU	1 GUU	1 GUG	1 GUC	1 GCA
Firmicutes	2 GAA	2 GUA	2 GAU	3 GUU	1 GUG	3 GUC	1 GCA
Cyanobacteria	1 GAA	1 GUA	2 GAU	1 GUU	1 GUG	1 GUC	1 GCA

Figure 6-3 Consensus tRNA abundances for two-fold degenerate amino acids with codons of the form NNY

Clade	Amino Acid: Two-fold degenerate Gln	Lys	Glu	Four-fold degenerate Pro	Thr	Val	Ala	Gly	Six-fold degenerate Leu	Ser	Arg
Alpha Pr	1 UUG 1 CUG	1 UUU 1 CUU	1 UUC 2 CUC	1 GGG 1 UGG 1 CGG	1 GGU 1 UGU 1 CGU	1 GAC 1 UAC 1 CAC	1 GGC 2 UGC 1 CGC	1 GCC 1 UCC 1 CCC	1 UAA 1 CAA 1 GAG 1 UAG 1 CAG	1 GCU 1 GGA 1 UGA 1 CGA	1 UCU 1 CCU 1 ACG 1 CCG
Gamma Pr	3 UUG	3 UUU	4 UUC	1 GGG 3 UGG	1 GGU 2 UGU	1 GAC 4 UAC	1 GGC 4 UGC	4 GCC 1 UCC	2 UAA 1 CAA 1 GAG 3 UAG 1 CAG	1 GCU 1 GGA 2 UGA	1 UCU 1 CCU 4 ACG 1 CCG
Xanthomonales (Gamma Pr)	1 UUG 1 CUG	1 UUU 1 CUU	1 UUC 2 CUC	1 GGG 1 UGG 1 CGG	1 GGU 1 UGU 1 CGU	1 GAC 1 UAC 1 CAC	1 GGC 2 UGC 1 CGC	2 GCC 1 UCC 1 CCC	1 UAA 1 CAA 1 GAG 1 UAG 2 CAG	1 GCU 1 GGA 1 UGA 1 CGA	1 UCU 1 CCU 2 ACG 1 CCG
Escol, Salent (Gamma Pr)	2 UUG 2 CUG	6 UUU	4 UUC	1 GGG 1 UGG 1 CGG	2 GGU 1 UGU 2 CGU	2 GAC 5 UAC	2 GGC 3 UGC	4 GCC 1 UCC 1 CCC	1 UAA 1 CAA 1 GAG 1 UAG 4 CAG	1 GCU 2 GGA 1 UGA 1 CGA	2 UCU 2 CCU 4 ACG 1 CCG
Actinobacteria	1 UUG 1 CUG	1 UUU 1 CUU	1 UUC 2 CUC	1 GGG 1 UGG 1 CGG	1 GGU 1 UGU 1 CGU	1 GAC 1 UAC 1 CAC	1 GGC 1 UGC 1 CGC	2 GCC 1 UCC 1 CCC	1 UAA 1 CAA 1 GAG 1 UAG 1 CAG	1 GCU 1 GGA 1 UGA 1 CGA	1 UCU 1 CCU 1 ACG 1 CCG
Firmicutes	3 UUG	3 UUU	4 UUC	2 UGG	1 GGU 3 UGU	4 UAC	4 UGC	3 GCC 3 UCC	2 UAA 1 CAA 1 GAG 2 UAG	1 GCU 1 GGA 2 UGA	1 UCU 1 CCU 2 ACG 1 CCG
Cyanobacteria	1 UUG	1 UUU	1 UUC	1 GGG 1 UGG 1 CGG	1 GGU 1 UGU 1 CGU	1 GAC 1 UAC	1 GGC 1 UGC 1 CGC	1 GCC 1 UCC 1 CCC	1 UAA 1 CAA 1 GAG 1 UAG 1 CAG	1 GCU 1 GGA 1 UGA 1 CGA	1 UCU 1 CCU 1 ACG 1 CCG

Figure 6-4 Consensus tRNA abundances for the remaining amino acids

copy number. The tRNA species with anticodons of the form GNN were seen for threonine and glycine but the tRNA species with anticodons of the form CNN was not seen in the Firmicutes.

6.3.2.3 Alpha Proteobacteria

For the two-fold degenerate amino acids with codons of the form NNR (Gln, Lys and Glu) the Alpha Proteobacteria were seen (as discussed in chapter 5) to prefer codons of the form NNG over the NNA alterative in the case of glutamine and lysine, but not for glutamate where either no preference was shown or NNA was preferred over NNG. The tRNA consensus (taken from tRNA gene copy numbers in the Alpha Proteobacteria) showed the presence of both tRNA species with anticodon UNN and with anticodon CNN (Figure 6-4) for all three of these amino acids. These tRNA species were present in low copy number for each amino acid in the Alpha Proteobacterial clade. This was in constrast to clades preferring just NNA codons where the only tRNA species had anticodons of the form UNN. The tRNA abundance data supports the use of the NNG codon for lysine and glutamine (i.e. CNN anticodon present) and tRNA abundances would equally support the use of the NNG codon for glutamate, but no switch in codon usage was observed for this amino acid.

The four-fold degenerate amino acids proline and threonine were seen to show CCG and ACC preference, respectively, in the Alpha Proteobacteria. The tRNA consenus for the Alpha Proteobacteria showed the presence of GNN, UNN and CNN anticodons in single copy number for these two amino acids allowing the CCG and ACC codon preferences to be translated by perfect match (CCG:CGG and ACC:GGU). However any codon usage would be possible given the tRNA species present in the Alpha Proteobacteria. The remaining four-fold degenerate amino acids were seen to use codons of the form NNU or NNU and NNA in the previous chapter. However, the tRNA species present for these amino acids was the same for valine, alanine and glycine as for threonine and proline, but no switch in codon usage was seen for these amino acids.

6.3.2.4 *Xanthomonas* species of the Gamma Proteobacteria

The G+C rich *Xanthomonas* species of the Gamma Proteobacteria were seen to have similar patterns of codon usage to the Alpha Proteobacteria, preferring NNG codons for the amino acids glutamine, lysine and proline and ACC for threonine. The tRNA abundances for the *Xanthomonas* species were also seen to be the same as those of the Alpha Proteobacteria for these amino acids with tRNA species with anticodons of the form CNN present.

6.3.2.5 *E. coli* and *S. enterica* of the Gamma Proteobacteria

These two Gamma Proteobacterial genomes were seen in the previous chapter to have similar patterns of codon usage to the main Gamma Proteobacterial clade except for the amino acids glutamine and proline (and possibly leucine) where codons of the form NNG were preferred instead of the NNA preference in the other Gamma Proteobacteria. The tRNA gene complements for these two bacteria were similar to those of the other Gamma Proteobacteria for the amino acids lysine, glutamate, valine, glycine and arginine. However, for the amino acids glutamine, proline, threonine, glycine and serine tRNA species with anticodons of the form CNN were additionally present. Changes in tRNA abundance therefore seem to be correlated with the changes in codon usage for glutamine, proline and possibly leucine but similar changes in codon usage for threonine did not initiate a switch in codon preference.

6.3.2.6 Actinobacteria

The Actinobacteria are G+C rich and as such were seen to show preference for NNG codons when, in the previous chapter, codon preference in two-fold degenerate amino acids with codons of the form NNR was considered. The consensus tRNA gene complements for these three amino acids (glutamine, lysine and glutamate) in the Actinobacteria showed that both tRNA species with anticodon UNN and CNN were used and present in low copy number.

The Actinobacteria also showed preference for CCG and ACC codons of the amino acids proline and lysine respectively and a general NNU or NNU and NNA codon preference for the other four-fold degenerate amino acids. The consensus tRNA abundances for the Actinobacteria showed the presence of the tRNA species with anticodon CNN along with UNN and GNN anticodon species for all of the four-fold degenerate amino acids.

6.3.2.7 Cyanobacteria

When codon usage for this clade was examined the only clear switch in codon usage observed was for the amino acid threonine. This amino acid showed ACC codon preference and whilst other amino acids showed signs of switches in codon preference the lack of optimal codons for many amino acids in many of the species meant other switches in codon usage could not be decided upon. The consensus tRNA gene complements for these bacteria were intriguing as the two-fold degenerate amino acids with codons of the form NNR showed the presence of only tRNA species with anticodons of the form UNN, similar to those seen in the Gamma Proteobacteria and Firmicutes. However, tRNA abundances for the four-fold degenerate amino acids showed tRNA abundances similar to those of the Actinobacteria and Alpha Proteobacteria with the presence of tRNA species with anticodons of the form GNN, UNN and CNN for the majority of four-fold degenerate amino acids (except valine where no CNN anticodon species was present).

6.3.3 Summarising the switching patterns

The examination of patterns of codon usage preference and tRNA abundances showed a strong correlation between the two. This was particularly pronounced for the two-fold degenerate amino acids glutamine, lysine and glutamate as well as the four-fold degenerate amino acids threonine and proline. For these amino acids clear switches in codon preference were seen as discussed in the previous chapter. It was now also evident that tRNA abundances reflect this change in codon preference. To make this clear switches in tRNA abundance were overlaid onto the bacterial phylogeny and can be seen to coincide with switches in codon preference (Figure 6-5).

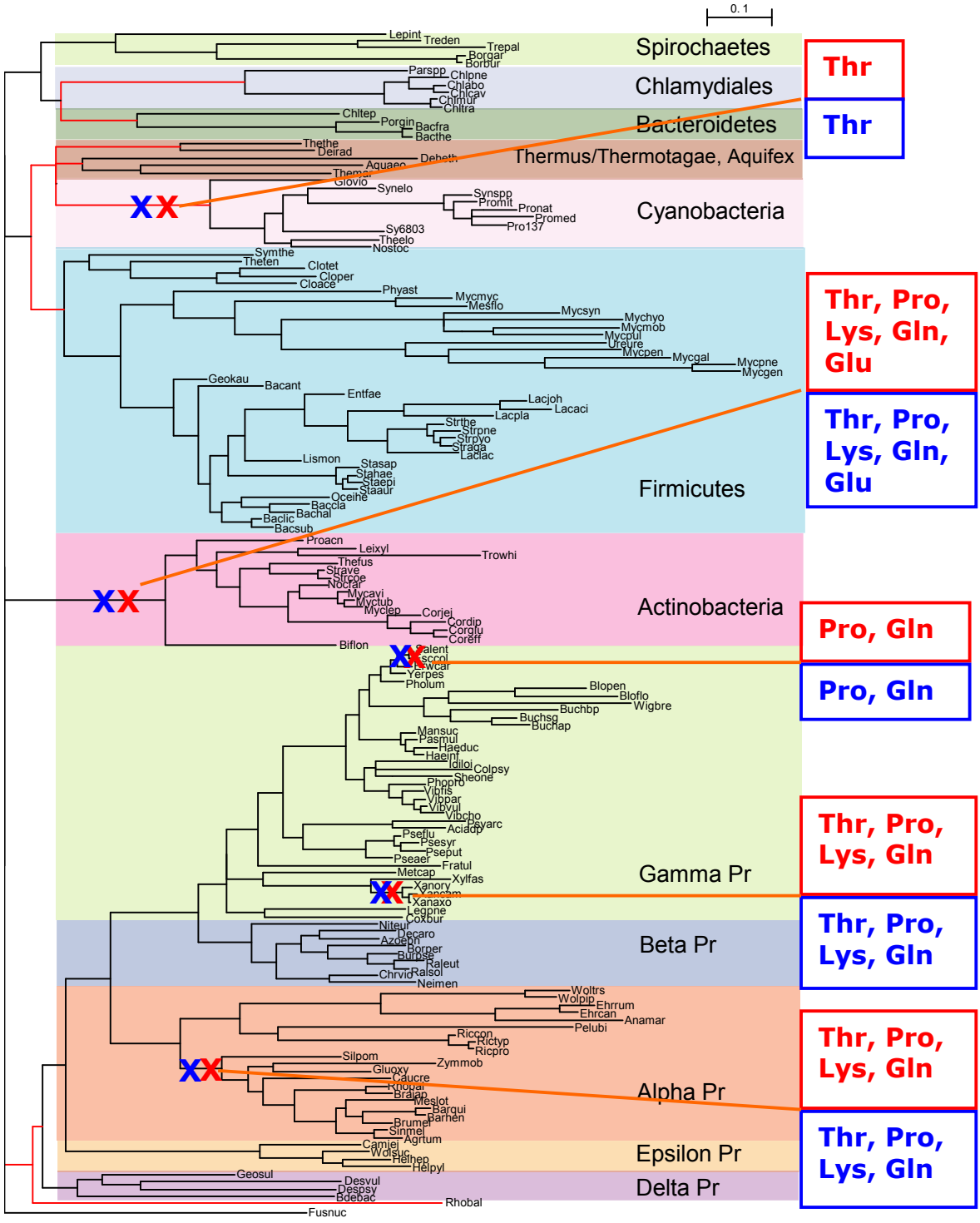

Figure 6-5 Phylogeny of the 160 bacterial species used in this study. Red crosses marking switches in codon preference towards G+C-ending codons. Blue crosses indicate the acquisition of a tRNA species with anticodon of the form CNN. The amino acids where relevant changes occur are labeled in boxes to the right.

The A+U rich clades such as the Firmicutes and the main Gamma Proteobacterial clade were seen to use codons of the form NNA for glutamine, glutamate, lysine, and proline as well as NNU codons for threonine. The tRNA abundances examined showed sole the use of tRNA species with anticodons of the form UNN (complementary to the NNA codon preferred) for glutamine, lysine and glutamate. For the amino acids proline and threonine tRNA species with anticodons of the form UNN were always the most abundant (present in multiple copies within the genome) and GNN anticodons in low copy number were also often seen but the CNN anticodon was not present.

The G+C rich genomes of the Alpha Proteobacteria, Actinobacteria and *Xanthomonas* species of the Gamma proteobacteria all showed alternative codon usage. Codons of the form NNG were seen to be preferred for the amino acids glutamine, lysine and proline (also glutamate in the Actinobacteria) whilst the ACC codon was preferred for threonine. In all cases such a switch in codon usage was seen to result in the acquisition of tRNA species with anticodons of the form CNN, although tRNA species with anticodons of the form UNN (and GNN for proline and threonine) were retained but in lower copy number than seen for the A+U rich species. In general G+C rich genomes were seen to have a greater variety of tRNA species available but in lower copy number than the A+U rich genomes where tRNA species with UNN anticodons were often present in multiple copies but no tRNA species with CNN anticodons were present.

Although changes in tRNA abundances were seen to have occurred when switches in codon preference occurred, it was also noticeable that on some occasions changes in tRNA abundances were not accompanied by switches in optimal codon choice. This was particularly true for the amino acids valine, alanine and glycine where tRNA abundances were similar to those of threonine and proline but no clear switch in codon preference was seen.

6.4 Discussion

6.4.1 Differences in codon and tRNA switching between amino acids

In the previous chapter it was noted that individual amino acids showed specific codon switching patterns. Additionally previous work looking at codon:anticodon preferences proposed three models to explain codon and tRNA choice (outlined in Rocha, 2004). These models were the *frequency model*, the *perfect match model* and the *stability model* (as discussed in the introduction to this chapter). The first two of these models appear to fit the data here under certain conditions.

Firstly, the amino acids phenylalanine, tyrosine, isoleucine, asparagine, histidine, aspartate and to a lesser extent the rare amino acid cysteine all use the NNC codon over the NNU codon when subject to selection on their codon usage. The basis of this is thought to be due to the presence of just a single tRNA species for each of the amino acids which always take the form GNN and therefore pairs exactly with the NNC codon ensuring fast and efficient translation of highly expressed genes within the genome in question. This form of codon:anticodon relationship appears to be an excellent example of the perfect match model as suggested by Ikemura (Ikemura, 1981a; Ikemura, 1981b). The GNN anticodon is also able to translate the NNU codon through wobble enabling non-optimal codons to also be translated but with less accuracy and efficiency than the preferred NNC codons. Anticodons with A at their first position appear to be more or less uniformly avoided; this may be related to the stereochemical destabilization of the codon:anticodon duplex although this is by no means certain. The avoidance of anticodons of the form ANN is universal for these amino acids with codons of the form NNY and so the NNU codon is never the preferred optimal codon and no switching takes place.

The remaining amino acids with only two synonymous codons (i.e. glutamine, lysine and glutamate) have codons that take the form NNR. These amino acids are seen to be influenced by directional selection since

they switch between preference for NNA codons or codons of the form NNG. This change indeed seems to be reflected in the tRNA abundances related to these amino acids although it is noticeable that genomes preferring codons of the form NNG have tRNAs with both CNN and UNN anticodons whilst those preferring codons of the form NNA largely just have tRNAs with UNN anticodons. The switch in codon preference between NNA and NNG is very clear and again it appears that tRNA species that pair with the codon exactly are preferred, at least for the case of the species preferring NNA codons. This necessity for the use of exact binding between codon and anticodon may again be a result of the two-fold degeneneracy seen for these amino acids as tRNA modification introduces the chance that incorrect amino acids could be included in the polypeptide chain (see Table 6-1 and the introduction to this chapter). This again appears to support the perfect match model of codon:anticodon recognition. In addition to this one can see that although some (usually A+T rich) genomes have solely UNN anticodons, G+C rich genomes have tRNAs with UNN and CNN anticodon species for these three amino acids. This is because an unmodified CNN anticodon only pairs with codons of the form NNG but UNN anticodons can pair with both NNA through Watson-Crick pairing and NNG through wobble (Table 6-1). These tRNA species with UNN anticodons are therefore required for the translation of non-optimal codons for these amino acids in G+C rich genomes as the optimal, CNN, tRNA species cannot usually pair with these codons.

For the amino acids each encoded by four synonymous codons there again appear to be two major groups, as noted in the previous chapter. The amino acids proline and threonine are susceptible to directional selection while valine, alanine and glycine show a predominant preference for synonymous codons of the form GBU. There is also a major distinction even between proline and threonine, with proline codon switching being primarily between the synonymous codons CCA and CCG, whereas threonine instead uses either ACU or ACC. Why this distinction is seen remains unclear but preferred proline codons tend to take on the form YYR whilst threonine codons take the form RYY. In the case of both these amino acids G+C rich species tend to have tRNA species with anticodons of

the form GNN, UNN and CNN whilst A+T rich species have GNN and UNN or even just UNN species. G+C rich species with all three tRNA species once again had these tRNAs in low copy number within the genome whilst A+T rich species often had tRNA species with UNN anticodons present in multiple copies within the genome.

The remaining four-fold degenerate amino acids (valine, alanine and glycine) use optimal codons of the form NNU which at first seems counter intuitive as these amino acids have no tRNA species with anticodons of the form ANN. It seems however that these four fold degenerate amino acids do not have the same restrictions as two fold degenerate amino acids where pairing by wobble can cause the incorporation of incorrect amino acids (especially when considering modified anticodons, see section 6.4.2), therefore it is possible that using NNU codons allow the maximization of the tRNA pool available by being able to use both GNN and UNN tRNA species. This model of codon:anticodon choice is consistent with the frequency model hypothesis which takes into account that anticodons can read several codons and that a codon be read by several different anticodons (Rocha, 2004). In the frequency model the most frequent codon is that that can be decoded by the largest number of tRNAs. When considering four-fold degenerate amino acids the problem of incorporating an incorrect amino acid through incorrect base pairing at the third codon position is not present; this means that the perfect match problem is less relevant, as appears to be the case for the amino acids valine, alanine and glycine. The ability to pair through wobble also possibly explains why the majority of genomes retain both GNN and UNN tRNAs but CNN is lost or gained under the influence of genomic G+C content. This is because a tRNA with anticodon of the form UNN can recognize most cognate amino acids as U, or some modified nucleosides derived from U, can pair with all synonymous codons (Table 6-1). Additionally a tRNA with an anticodon of the form GNN can recognize both NNC and NNU anticodons. In contrast, CNN anticodons are usually specific to NNG codons and so have limited use when NNG codons are not optimal.

If the frequency model hypothesis is true it does not explain why proline and threonine amino acids do not show a similar strategy, but it does explain the distinction between these four-fold degenerate amino acids and two-fold degenerate amino acids. The frequency model also appears to hold to some extent for six fold degenerate amino acids that also quite often use NNU codons and have a large number of different tRNA species for an individual amino acid. However, the situation for six-fold degenerate amino acids appears to be more complex and clearer patterns are harder to distinguish.

6.4.2 Relating changes in tRNA abundances to Shields' model

In the previous chapter the Shields' predictions for codon switching were discussed. It was seen that switches in codon usage are correlated with changes in genomic G+C content, which supports the model of codon switching proposed by Shields. In this chapter it was seen that where switches in optimal codon usage have taken place tRNA abundances (i.e. tRNA gene copy) were seen to change also. However, there were some occaisions where tRNA abundances were seen to change but no corresponding switch in codon usage was observed. In addition when looking at closely related species fluctuations in tRNA abundances can be seen and so it seems that tRNA gene complements are quite fluid but retain 'core' tRNAs whilst new tRNAs are lost and gained (See Appendix D for full tRNA gene complements). A changing mutational bias may cause these 'core' tRNA genes to change to reflect the new mutational bias whilst codon usage also switches. Such a mechanism explains the majority of switches in codon usage and tRNA gene complements (Alpha Proteobacteria, Actinobacteria and *Xanthomonas* species) and fits well with the predicts of Shields (1990).

The switch, for *E. coli* and *S. enterica* genomes, does not seem to be correlated with G+C content as these two genomes have a genomic GC3s of around 0.50. In this case the acquisition of new glutamine and proline tRNA species with anticodons of the form CNN seems to have allowed a

switch in codon usage (towards NNG codons) to take place, although the exact mechanism behind this is unclear.

6.5 Conclusions

In this chapter it was seen that tRNA abundances do indeed change in conjunction with switches in optimal codon usage. Where a switch in codon usage towards codons of the form NNG/C from codons of the form NNA/U was observed tRNA species with anticodon of the form CNN were seen to have been acquired. It was also seen that when the tRNA species of anticodon CNN was acquired the multiple copies of the tRNA species of anticodon UNN seemed to be lost leaving mainly single copies of each gene for particular tRNA species in the genome. Fluctuations in tRNA gene complement were seen in closely related species and there is additional research showing that even different strains of the same species can have different tRNA copy numbers (Withers *et al.,* 2006). This indicates that tRNA evolution may be a lot less restricted than previously thought with tRNA gene complements fluctuating around a core set of tRNA genes and those core genes changing as mutational bias changes and codon usage switches.

Chapter 7:
Conclusions and Future Directions

7.1 Conclusions

The majority of work concerning codon usage in bacteria has previously been concerned with patterns of codon usage within individual bacterial genomes. Some bacterial species have been seen to have codon usage heavily influenced by selection with the classical example being that of *Escherichia coli*. Codon usage of the Delta Proteobacterium *Bdellovibrio bacteriovorus* was investigated in chapter three using correspondence analysis and it was seen that translational selection was indeed the primary factor influencing codon usage in this genome also. It is often assumed that all bacterial species have codon usage influenced by translational selection but bacterial species such as *Helicobacter pylori* show little evidence of selected codon usage bias with the effects of neutral mutation being more evident instead. It seems that pattern of codon usage within bacterial genomes seems to be a balance between the effects of selection on one hand and mutation combined with drift on the other.

The two key problems in looking at the strength of selected codon usage bias across many bacterial genomes are the huge variation in genomic G+C content in bacterial genomes and the fact that optimal codon choice varies between bacterial genomes. Previous work by Sharp *et al.* (2005) aimed to overcome these problems and quantify the strength of selected codon usage bias across bacterial genomes using the S_{WWY} statistic. Firstly, the statistic was designed to be unaffected by variation in optimal codon choice by using four amino acids where optimal codon choice was conserved across all bacteria. These four amino acids (phenylalanine, tyrosine, isoleucine and asparagine) had synonymous codons of the form WWY and it was expected that the WWC codon should always be the preferred optimal codon over the WWU alternative as the only isoaccepting tRNA species for these amino acids in bacterial species are tRNA species with anticodons of the form GWW. The WWC codon and GWW anticodon match perfectly by

Watson-Crick base pairing and so ensure accurate efficient translation. Secondly, the statistic was designed to be unaffected by genomic G+C content as the effects of mutational bias were limited by comparing the codon usage in highly expressed genes within a genome (genes expected to be influenced by selected codon usage bias) with the genome as a whole (largely uninfluenced by selected codon usage bias with codon usage more representative of the underlying mutational biases within the genome).

The study by Sharp *et al.* (2005) used the S_{WWY} statistic to investigate the strength of selected codon usage bias in 80 bacterial genomes and found varying levels of selection across these bacterial genomes. Bacteria such as *Clostridium perfringens* (S_{WWY} of 2.65) showed strong selected codon usage bias but others such as *Helicobacter pylori* (S_{WWY} of 0.02) showed little evidence of selected codon usage bias. The work presented in this thesis aimed, initially, to extend this original study to include newly sequenced bacterial genomes. In the new 160 genome dataset the bacteria represented were now even more diverse than previously seen with the addition of new groups such as the Delta Proteobacteria. More species were represented in each major clade allowing a much more comprehensive analysis of selected codon usage bias across bacterial genomes. The new species added were widely distributed across the original phylogeny but were particularly useful in adding more resolution to underrepresented clades. The Alpha Proteobacteria gained more resolution to its G+C-poor Rickettsiales clade, whilst the Beta Proteobacteria also increased from just three species to nine species. The Gamma Proteobacteria and Firmicutes were already well represented but were even more so in this new larger dataset. The Delta Proteobacteria had no representation at all in the original dataset but now had four species whilst the epsilon proteobacteria doubled in size from two to four represented genomes. The resolution of the Actinobacteria was also greatly enhanced with the addition of seven new genomes to take the total to 15 genomes. The strength of selected codon usage bias was again seen to vary, with 66% (105/160) of species exhibiting a significant amount of selection. In chapter four of this thesis the strength of selected codon usage was shown to be highly correlated with (the log of) generation time. This had already been inferred by Sharp

et al. (2005) using rRNA operon number but was done directly here using generation time data. Bacteria with a rapid generation time were seen to be more heavily influenced by selected codon usage bias than slow growing bacteria whose highly expressed genes were less constrained in their codon usage. This may be expected as bacteria with a rapid generation time should require faster and more efficient translation and, therefore, have codon usage in highly expressed genes that is highly co-adapted to the most abundant tRNA species present to ensure the optimization of the translational machinery. Indeed the bacterial species with the highest S_{WWY} value (*Clostridium Perfringens* with an S_{WWY} of 2.65), indicating the strongest degree of selected codon usage bias of all the 160 bacterial species, had a generation time in the order of just a few minutes. In contrast, bacterial species with low S_{WWY} values showed little requirement for highly co-adapted codon usage and tRNA abundances with generation times in the order of days.

Although codon usage in the four amino acids (Phe, Tyr, Ile, Asn) used to calculate the strength of selected codon usage bias (S_{WWY}) showed conserved optimal codon usage across all bacterial genomes this was not the case for all amino acids. How such highly co-adapted optimal codon usage and tRNA abundances could change was not well understood. It seemed that a relaxation in selection must have had to occur followed by genetic drift and a reinstatement of selection in order for these changes in optimal codon preference to occur. If selection were maintained one would expect that a change in either optimal codon usage or tRNA abundances would disrupt the highly co-adapted translational machinery and, therefore, be highly disadvantageous and selected against. However, work by Shields (1990) provided a model whereby optimal codon usage and tRNA abundances could change without a relaxation of selection. This work predicted that sudden switches in optimal codon usage were possible in genomes under the influence of selected codon usage bias without a relaxation of selection if the change in mutational bias was strong enough. Shields suggested that genes under weak selection should change their codon usage in line with the change in mutational bias whilst genes under strong selection should maintain their optimal codon usage against the

change in mutational bias until the bias was so strong that the current optimal codon usage would become unsustainable. Such a situation was predicted to arise when the new G+C content of the genome resulted in a situation whereby the codon usage of the genome as a whole was not well served by the tRNA gene complement. Due to the highly co-adapted nature of optimal codon usage and tRNA abundances the tRNA gene complement would first be expected to change thus changing the identity of the optimal codons, therby exerting selection on highly expressed genes within the genome to change their codon usage to use these new optimal codons.

In order to investigate optimal codon choice for other amino acids the S_{WWY} statistic was modified. Where bacteria showed significant levels of selected codon usage bias optimal codon choice in highly expressed genes was compared for each amino acid among bacteria. By plotting the strength of selected codon usage bias for each amino acid, S_{aa}, against the S_{WWY} statistic calculated for phenyalanine, tyrosine, isoleucine and asparagine switches in codon preference in bacteria were looked for in other two-fold degenerate amino acids. Amino acids with codons of the form NNY (His and Asp) were seen to show similar optimal codon choice to the four amino acids with codons of the form WWY, where the NNC codon was always preferred over the NNU codon. However, for two-fold degenerate amino acids with codons of the form NNR (Gln, Lys and Glu) switches in codon usage were evident with optimal codon choice separating the bacterial species into two groups based on clade and corresponding genomic G+C content. Bacterial species with with high genomic G+C content and highly expressed genes under the influence of selected codon usage bias, such as the Alpha Proteobacteria, the Actinobacteria and *Xanthomonas* species of the Gamma Proteobacteria, largely showed a preference for NNG codons over the NNA alternative. In contrast A+U rich bacterial species such as the Firmicutes and the main clade of the Gamma Proteobacteria showed NNA codon preference. For some of the four-fold degenerate amino acids similar divisions were evident in patterns of optimal codon choice with the G+C rich clades preferring CCG codons for proline and ACC codons for threonine and the A+U rich clades preferring CCA codons for proline and ACU codons for threonine. In other four-fold degenerate amino acids less

evidence of codon switching was evident with codons of the form NNU uniformly preferred.

When tRNA abundances were investigated it was evident that switches in codon usage were correlated with changes in tRNA abundances (as inferred by changes in tRNA gene complement). For two-fold degenerate amino acids of the form NNR (Gln, Lys and Glu) a switch in codon usage from NNA in the Firmicutes and Gamma Proteobacteria to NNG in the Alpha Proteobacteria, Actinobacteria and *Xanthomonas* species was correlated with the acquisition of tRNA species with anticodon of the form CNN (switch in codon usage for glutamate to NNG and corresponding acquisition of the CNN anticodon was only seen in Actinobacteria). Similar patterns of codon usage and correlated tRNA abundances were seen for the amino acids proline and threonine. Corresponding switches in optimal codon usage and tRNA gene complement were seen in bacterial genomes under selection and appeared to be influenced by changes in mutational bias as indicated by genomic G+C content. These observations demonstrated that for the amino acids glutamine, lysine, glutamate, proline and threonine changes in optimal codon preference and tRNA abundances seem to be correlated with changes in mutational bias. Such observations are as one would expect if the model proposed by Shields (1990) were true.

7.2 Future Directions

The four amino acids phenylalanine, tyrosine, isoleucine and asparagine were used to calculate the strength of selected codon usage bias, S_{WWY}, as they showed conserved optimal codon choice across all bacterial genomes. From the work in this thesis it appears that the other amino acids with codons of the form NNY show similar patterns of codon usage (apart from the often rare amino acid cysteine). It would, therefore, make sense to extend the calculation of the strength of selected codon usage bias statistic to include the amino acids histidine and aspartate in the future.

Another immediately advantageous way to further the work here would be to once again increase the size of the dataset. Bacterial genomes are being

sequenced at a rapid rate and so many more bacterial species are now available for analysis. Such an expansion of the dataset should allow patterns of codon usage to be explored in clades underrepresented in this dataset. A greater number of Cyanobacterial or Beta Proteobacterial species would be useful in establishing what patterns of selected codon usage bias are present in such species, if indeed there is a significant level of selected codon usage bias at all. Other underrepresented clades such as the Spirochaetes, Chlamydiales or Bacteroidetes would also benefit from such and expansion of the dataset, hopefully allowing clearer patterns of codon usage to emerge for these clades.

The examination of patterns of optimal codon usage and corresponding tRNA abundances could also be widened into other unicellular prokaryotes such as the archaea. It would be interesting to see if the archaea have many species with codon usage influenced by selected codon usage bias and whether similar codon switching patterns, influenced by genomic G+C content, were visible in the archaea. A preliminary scan of the fully sequence archaeal genomes available shows that some Methanobacteriales, Methanococcales and Methanosarcinales have as many as 4 rRNA operons and 40-60 tRNA genes, indicating that many of these genomes may have codon usage patterns influenced by selected codon usage bias. Work by McInerney (1997) analyzing patterns of codon usage bias in *Methanocaldococcus jannaschii* also showed evidence of translational selection in this archaeal genome. Additionally, many archaeal genomes such as the *Methanococci* are A+U rich (genomic G+C content of around 0.33) but some archaea seem to have G+C contents as high as 0.67, for example among the *Halobacteria*, and so patterns of optimal codon switching corresponding to changes in mutational bias are possible (but only if selected codon usage bias is detected in these species).

Further work could also be done to look at how changes in tRNA gene complement occur in bacterial genomes with regard to the changes in mutational bias and codon usage within bacterial species. These switches in codon preference between A+U rich bacteria (using A- and U-ending

codons) to G+C rich bacteria (using G- and C-ending codons) were accompanied by the acquisition of tRNA species with the anticodon CNN, along with the loss of many UNN gene copies, in many G+C rich genomes. How such events have occurred warrants further investigation. In particular, it would be interesting to know how these changes in tRNA genes occur. It is conceivable that these new tRNA species evolve by mutation in the anticodon of a UNN gene, or by horizontal transfer of a CNN gene from another bacterium. Extra tRNA genes may also arise by local duplication events and tRNA loss by local deletion. To further examine some of these ideas one could take tRNA gene sequences for a particular amino acid (for example glutamine) from species in the Alpha Proteobacteria, Actinobacteria, Firmicutes and Gamma Proteobacteria; where the latter two have only tRNA species with UUG anticodons whilst the former have both UUG and CUG anticodons. One could then see if the CUG tRNAs from different clades cluster phylogenetically or whether CUG and UUG tRNAs cluster with those from the same clade. If the CUG tRNAs from different clades clustered togther then the indication would be that the CUG tRNA species was acquired by horizontal gene transfer. However, if CUG and UUG tRNAs clustered with those from the same clade gene duplication and/or mutation would be more likely. Therefore, such a method would allow some of the questions regarding the methods of tRNA evolution to be addressed.

The rate at which genomic information is being acquired is increasing at an almost exponential rate. This increase in data available should, over time, allow patterns of codon usage in prokaryotes to become better and better understood. As this happens techniques, such as those used in this thesis, will become evermore important in understanding this wealth of sequence data available.

Bibliography

Agris, P. F. (2004). Decoding the genome: a modified view. *Nucleic Acids Res* **32**, 223-238.

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

Akashi, H. & Eyre-Walker, A. (1998). Translational selection and molecular evolution. *Curr Opin Genet Dev* **8**, 688-693.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol* **215**, 403-410.

Andersson, S. G. & Kurland, C. G. (1990). Codon preferences in free-living microorganisms. *Microbiol Rev* **54**, 198-210.

Andersson, S. G. & Sharp, P. M. (1996). Codon usage and base composition in *Rickettsia prowazekii*. *J Mol Evol* **42**, 525-536.

Andersson, J. O. & Andersson, S. G. (1999). Genome degradation is an ongoing process in *Rickettsia*. *Mol Biol Evol* **16**, 1178-1191.

Barel, G. & Jurkevitch, E. (2001). Analysis of phenotypic diversity among host-independent mutants of *Bdellovibrio bacteriovorus* 109J. *Arch Microbiol* **176**, 211-216.

Battistuzzi, F. U., Feijao, A. & Hedges, S. B. (2004). A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land. *BMC Evol Biol* **4**, 44.

Beck, S., Schwudke, D., Strauch, E., Appel, B. & Linscheid, M. (2004). *Bdellovibrio bacteriovorus* strains produce a novel major outer

membrane protein during predacious growth in the periplasm of prey bacteria. *J Bacteriol* **186**, 2766-2773.

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

Benzécri, J.-P. (1983). Analyse de l'inertie intra-classe par l'analyse d'un tableau des correspondances. *Les Cahiers de l'Analyse des Données* **8**, 351-358.

Bern, M. & Goldberg, D. (2005). Automatic selection of representative proteins for bacterial phylogeny. *BMC Evol Biol* **5**, 34.

Bern, M., Goldberg, D. & Lyashenko, E. (2006). Data mining for proteins characteristic of clades. *Nucleic Acids Res* **34**, 4342-4353.

Bernardi, G. & Bernardi, G. (1986). Compositional constraints and genome evolution. *J Mol Evol* **24**, 1-11.

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

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

Ciccarelli, F. D., Doerks, T., von Mering, C., Creevey, C. J., Snel, B. & Bork, P. (2006). Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**, 1283-1287.

Cochella, L. & Green, R. (2004). Wobble during decoding: more than third-position promiscuity. *Nat Struct Mol Biol* **11**, 1160-1162.

Cox, E. C. & Yanofsky, C. (1967). Altered base ratios in the DNA of an *Escherichia coli* mutator strain. *Proc Natl Acad Sci U S A* **58**, 1895-1902.

Crick, F. H. (1966). Codon--anticodon pairing: the wobble hypothesis. *J Mol Biol* **19**, 548-555.

Das, S., Paul, S., Chatterjee, S. & Dutta, C. (2005). Codon and amino Acid usage in two major human pathogens of genus bartonella -- optimization between replicational-transcriptional selection, translational control and cost minimization. *DNA Res* **12**, 91-102.

Davies, J. (1996). Origins and evolution of antibiotic resistance. *Microbiologia* **12**, 9-16.

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

Dufayard, J. F., Duret, L., Penel, S., Gouy, M., Rechenmann, F. & Perrière, G. (2005). Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics* **21**, 2596-2603.

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

Fichant, G. A. & Burks, C. (1991). Identifying potential tRNA genes in genomic DNA sequences. *J Mol Biol* **220**, 659-671.

Francino, M. P. & Ochman, H. (1997). Strand asymmetries in DNA evolution. *Trends Genet* **13**, 240-245.

Francino, M. P. & Ochman, H. (2001). Deamination as the basis of strand-asymmetric evolution in transcribed *Escherichia coli* sequences. *Mol Biol Evol* **18**, 1147-1150.

Galtier, N. & Lobry, J. R. (1997). Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. *J Mol Evol* **44**, 632-636.

Galtier, N., Tourasse, N. & Gouy, M. (1999). A nonhyperthermophilic common ancestor to extant life forms. *Science* **283**, 220-221.

Gouy, M. & Grantham, R. (1980). Polypeptide elongation and tRNA cycling in *Escherichia coli*: a dynamic approach. *FEBS Lett* **115**, 151-155.

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

Gouy, M., Gautier, C., Attimonelli, M., Lanave, C. & di Paola, G. (1985). ACNUC--a portable retrieval system for nucleic acid sequence databases: logical and physical designs and usage. *Comput Appl Biosci* **1**, 167-172.

Grantham, R., Gautier, C. & Gouy, M. (1980a). Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type. *Nucleic Acids Res* **8**, 1893-1912.

Grantham, R., Gautier, C., Gouy, M., Mercier, R. & Pave, A. (1980b). Codon catalog usage and the genome hypothesis. *Nucleic Acids Res* **8**, r49-r62.

Grantham, R., Gautier, C., Gouy, M., Jacobzone, M. & Mercier, R. (1981). Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Res* **9**, r43-74.

Greenacre, M. J. (1984). *Theory and Applications of Correspondence Analysis*: Academic Press.

Grocock, R. J. & Sharp, P. M. (2001). Synonymous codon usage in *Cryptosporidium parvum*: identification of two distinct trends among genes. *Int J Parasitol* **31**, 402-412.

Grocock, R. J. & Sharp, P. M. (2002). Synonymous codon usage in *Pseudomonas aeruginosa* PA01. *Gene* **289**, 131-139.

Grosjean, H. & Fiers, W. (1982). Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes. *Gene* **18**, 199-209.

Hartl, D. L., Moriyama, E. N. & Sawyer, S. A. (1994). Selection intensity for codon bias. *Genetics* **138**, 227-234.

Haubold, B. & Wiehe, T. (2004). Comparative genomics: methods and applications. *Naturwissenschaften* **91**, 405-421.

Henry, I. & Sharp, P. M. (2007). Predicting gene expression level from codon usage bias. *Mol Biol Evol* **24**, 10-12.

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

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

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

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

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**, 143-155.

Klappenbach, J. A., Dunbar, J. M. & Schmidt, T. M. (2000). rRNA operon copy number reflects ecological strategies of bacteria. *Appl Environ Microbiol* **66**, 1328-1333.

Kobayashi, I. (2001). Behavior of restriction-modification systems as selfish mobile elements and their impact on genome evolution. *Nucleic Acids Res* **29**, 3742-3756.

Konigsberg, W. & Godson, G. N. (1983). Evidence for use of rare codons in the *dnaG* gene and other regulatory genes of *Escherichia coli*. *Proc Natl Acad Sci U S A* **80**, 687-691.

Kurland, C. G. (1991). Codon bias and gene expression. *FEBS Lett* **285**, 165-169.

Kyte, J. & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J Mol Biol* **157**, 105-132.

Lafay, B., Atherton, J. C. & Sharp, P. M. (2000). Absence of translationally selected synonymous codon usage bias in *Helicobacter pylori*. *Microbiology* **146**, 851-860.

Lawrence, J. G. & Ochman, H. (1997). Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol* **44**, 383-397.

Lawrence, J. G. & Ochman, H. (1998). Molecular archaeology of the *Escherichia coli* genome. *Proc Natl Acad Sci U S A* **95**, 9413-9417.

Lenz, R. & Hespell, R. (1978). Attempts to grow *Bdellovibrios* micurgically injected into animal cells. *Arch Micro* **119**, 245-248.

Li, W. H. (1987). Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. *J Mol Evol* **24**, 337-345.

Liljas, A. (1999). Function is structure. *Science* **285**, 2077-2078

Lim, V. I. & Curran, J. F. (2001). Analysis of codon:anticodon interactions within the ribosome provides new insights into codon reading and the genetic code structure. *Rna* **7**, 942-957.

Liu, Q. (2006). Analysis of codon usage pattern in the radioresistant bacterium *Deinococcus radiodurans*. *Biosystems* **85**, 99-106.

Lobry, J. R. (1996). Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol Biol Evol* **13**, 660-665.

Lobry, J. R. & Chessel, D. (2003). Internal correspondence analysis of codon and amino-acid usage in thermophilic bacteria. *J Appl Genet* **44**, 235-261.

Lobry, J. R. & Necsulea, A. (2006). Synonymous codon usage and its potential link with optimal growth temperature in prokaryotes. *Gene* **22**, 22.

Lowe, T. M. & Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**, 955-964.

McInerney, J. O. (1997). Prokaryotic genome evolution as assessed by multivariate analysis of codon usage patterns**.** *Microbial and Comparative Genomics* **2**, 89-97

Marashi, S. A. & Ghalanbor, Z. (2004). Correlations between genomic GC levels and optimal growth temperatures are not 'robust'. *Biochem Biophys Res Commun* **325**, 381-383.

McLean, M. J., Wolfe, K. H. & Devine, K. M. (1998). Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. *J Mol Evol* **47**, 691-696.

Medigue, C., Rouxel, T., Vigier, P., Henaut, A. & Danchin, A. (1991). Evidence for horizontal gene transfer in *Escherichia coli* speciation. *J Mol Biol* **222**, 851-856.

Mira, A. & Moran, N. A. (2002). Estimating population size and transmission bottlenecks in maternally transmitted endosymbiotic bacteria. *Microb Ecol* **44**, 137-143.

Moran, N. A. (1996). Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proc Natl Acad Sci U S A* **93**, 2873-2878.

Mrazek, J. & Karlin, S. (1998). Strand compositional asymmetry in bacterial and large viral genomes. *Proc Natl Acad Sci U S A* **95**, 3720-3725.

Musto, H., Naya, H., Zavala, A., Romero, H., Alvarez-Valin, F. & Bernardi, G. (2004). Correlations between genomic GC levels and optimal growth temperatures in prokaryotes. *FEBS Lett* **573**, 73-77.

Musto, H., Naya, H., Zavala, A., Romero, H., Alvarez-Valin, F. & Bernardi, G. (2005). The correlation between genomic G+C and optimal growth temperature of prokaryotes is robust: a reply to Marashi and Ghalanbor. *Biochem Biophys Res Commun* **330**, 357-360.

Musto, H., Naya, H., Zavala, A., Romero, H., Alvarez-Valin, F. & Bernardi, G. (2006). Genomic GC level, optimal growth temperature, and genome size in prokaryotes. *Biochem Biophys Res Commun* **347**, 1-3.

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

Nakamura, Y., Itoh, T., Matsuda, H. & Gojobori, T. (2004). Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nat Genet* **36**, 760-766.

Novembre, J. A. (2002). Accounting for background nucleotide composition when measuring codon usage bias. *Mol Biol Evol* **19**, 1390-1394.

Nunez, M. E., Martin, M. O., Duong, L. K., Ly, E. & Spain, E. M. (2003). Investigations into the life cycle of the bacterial predator *Bdellovibrio bacteriovorus* 109J at an interface by atomic force microscopy. *Biophys J* **84**, 3379-3388.

Ochman, H., Lawrence, J. G. & Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299-304.

Ohama, T., Muto, A. & Osawa, S. (1990). Role of GC-biased mutation pressure on synonymous codon choice in *Micrococcus luteus*, a bacterium with a high genomic GC-content. *Nucleic Acids Res* **18**, 1565-1569.

Ohkubo, S., Muto, A., Kawauchi, Y., Yamao, F. & Osawa, S. (1987). The ribosomal protein gene cluster of *Mycoplasma capricolum*. *Mol Gen Genet* **210**, 314-322.

Olsen, G. J., Larsen, N. & Woese, C. R. (1991). The ribosomal RNA database project. *Nucleic Acids Res* **19 Suppl**, 2017-2021.

Olsen, G. J., Woese, C. R. & Overbeek, R. (1994). The winds of (evolutionary) change: breathing new life into microbiology. *J Bacteriol* **176**, 1-6.

Osawa, S., Jukes, T. H., Watanabe, K. & Muto, A. (1992). Recent evidence for evolution of the genetic code. *Microbiol Rev* **56**, 229-264.

Pagel, M. (1999). Inferring the historical patterns of biological evolution. *Nature* **401**, 877-884.

Peden, J. (1999). Analysis of codon usage, PhD Thesis. University of Nottingham.

Perrière, G. & Thioulouse, J. (2002). Use and misuse of correspondence analysis in codon usage studies. *Nucleic Acids Res* **30**, 4548-4555.

Picardeau, M., Lobry, J. R. & Hinnebusch, B. J. (2000). Analyzing DNA strand compositional asymmetry to identify candidate replication origins of *Borrelia burgdorferi* linear and circular plasmids. *Genome Res* **10**, 1594-1604.

Post, L. E., Strycharz, G. D., Nomura, M., Lewis, H. & Dennis, P. P. (1979). Nucleotide sequence of the ribosomal protein gene cluster adjacent to the gene for RNA polymerase subunit *β* in *Escherichia coli*. *Proc Natl Acad Sci U S A* **76**, 1697-1701.

Post, L. E. & Nomura, M. (1980). DNA sequences from the *str* operon of *Escherichia coli*. *J Biol Chem* **255**, 4660-4666.

Remm, M., Storm, C. E. & Sonnhammer, E. L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol* **314**, 1041-1052.

Rendulic, S., Jagtap, P., Rosinus, A. & other authors (2004). A predator unmasked: life cycle of *Bdellovibrio bacteriovorus* from a genomic perspective. *Science* **303**, 689-692.

Rice, W. R. (1989). Analyzing tables of statistical tests. *Evolution* **43**, 223-225.

Rivera, M. C., Jain, R., Moore, J. E. & Lake, J. A. (1998). Genomic evidence for two functionally distinct gene classes. *Proc Natl Acad Sci U S A* **95**, 6239-6244.

Rocha, E. P., Danchin, A. & Viari, A. (1999). Universal replication biases in bacteria. *Mol Microbiol* **32**, 11-16.

Rocha, E. P. (2004). Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. *Genome Res* **14**, 2279-2286.

Romero, H., Zavala, A. & Musto, H. (2000). Codon usage in *Chlamydia trachomatis* is the result of strand-specific mutational biases and a complex pattern of selective forces. *Nucleic Acids Res* **28**, 2084-2090.

Ronquist, F. & Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572-1574.

Santos, M. A., Moura, G., Massey, S. E. & Tuite, M. F. (2004). Driving change: the evolution of alternative genetic codes. *Trends Genet* **20**, 95-102.

Sharp, P. M. & Li, W. H. (1986a). Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for 'rare' codons. *Nucleic Acids Res* **14**, 7737-7749.

Sharp, P. M. & Li, W. H. (1986b). An evolutionary perspective on synonymous codon usage in unicellular organisms. *J Mol Evol* **24**, 28-38.

Sharp, P. M. & Li, W. H. (1987a). The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* **15**, 1281-1295.

Sharp, P. M. & Li, W. H. (1987b). The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. *Mol Biol Evol* **4**, 222-230.

Sharp, P. M. (1990). Processes of genome evolution reflected by base frequency differences among *Serratia marcescens* genes. *Mol Microbiol* **4**, 119-122.

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

Shields, D. C. (1990). Switches in species-specific codon preferences: the influence of mutation biases. *J Mol Evol* **31**, 71-80.

Shimizu, T., Ohtani, K., Hirakawa, H. & other authors (2002). Complete genome sequence of *Clostridium perfringens*, an anaerobic flesh-eater. *Proc Natl Acad Sci U S A* **99**, 996-1001.

Stoletzki, N. & Eyre-Walker, A. (2007). Synonymous Codon Usage in *Escherichia coli*: Selection for Translational Accuracy. *Mol Biol Evol* **24**, 374-381.

Stolp, H. & Starr, M. P. (1963). *Bdellovibrio Bacteriovorus* Gen. Et Sp. N., a Predatory, Ectoparasitic, and Bacteriolytic Microorganism. *Antonie Van Leeuwenhoek* **29**, 217-248.

Sueoka, N. (1962). On the genetic basis of variation and heterogeneity of DNA base composition. *Proc Natl Acad Sci U S A* **48**, 582-592.

Sueoka, N. (1988). Directional mutation pressure and neutral molecular evolution. *Proc Natl Acad Sci U S A* **85**, 2653-2657.

Sueoka, N. (1995). Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. *J Mol Evol* **40**, 318-325.

Thioulouse, J., Chessel, D., Doledec, S. & Olivier, J. M. (1997). ADE-4: A multivariate analysis and graphical display software. *Statistics and Computing* **7**, 75-83.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673-4680.

Wang, H. C., Susko, E. & Roger, A. J. (2006). On the correlation between genomic G+C content and optimal growth temperature in prokaryotes: data quality and confounding factors. *Biochem Biophys Res Commun* **342**, 681-684.

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

Wernegreen, J. J., Degnan, P. H., Lazarus, A. B., Palacios, C. & Bordenstein, S. R. (2003). Genome evolution in an insect cell: distinct features of an ant-bacterial partnership. *Biol Bull* **204**, 221-231.

Wolf, Y. I., Rogozin, I. B., Grishin, N. V., Tatusov, R. L. & Koonin, E. V. (2001). Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol Biol* **1**, 8.

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

Wright, S. (1977). Evolution and the genetics of populations, vol 3. Experimental results and evolutionary deductions: University of Chicago Press, Chicago.

Yarian, C., Townsend, H., Czestkowski, W., Sochacka, E., Malkiewicz, A. J., Guenther, R., Miskiewicz, A. & Agris, P. F. (2002). Accurate translation of the genetic code depends on tRNA modified nucleosides. *J Biol Chem* **277**, 16391-16395.

Appendices

Appendix A

The full gene lists showing genes significantly up regulated in *B. bacteriovorus* as compared to *E. coli* and vice versa.

Appendix B

Paper looking at methods of predicting highly expressed genes using codon usage bias measures.

Appendix C

Data used to produce the codon switching plots shown in chapter 5.

Appendix D

Full list of tRNA abundances for all 160 bacterial genomes as extracted from the tRNAscan-SE Genomic tRNA Database and discussed in chapter 6.

Appendix A

Table of genes found to be potentially up-regulated in *B. bacteriovorus* as compared to *E. coli*

Bdellovibrio Accession Number	*Bdellovibrio* Fop	Description	*E. coli* Accession Number	*E. coli* Fop
BX842654.TUF	0.872	translation elongation factor Tu	AE000410.TUFA	0.793
BX842655.PE159	0.852	DNA-binding protein HU-alpha	AE000150.HUPB	0.642
BX842648.PE189	0.847	30S ribosomal protein S18	AE000491.RPSR	0.642
BX842654.RPLQ	0.834	50S ribosomal protein L17	AE000407.RPLQ	0.614
BX842654.RPSE	0.794	ribosomal protein S5	AE000408.RPSE	0.676
BX842654.RPLN	0.788	ribosomal protein L14	AE000408.RPLN	0.623
BX842656.NDK	0.782	nucleoside diphosphate kinase	AE000338.NDK	0.634
BX842653.FKPA	0.749	peptidyl-prolyl cis-trans isomerase, FKBP-type	AE000410.FKPA	0.640
BX842648.MDH	0.748	malate dehydrogenase	AE000403.MDH	0.599
BX842646.RPME	0.736	ribosomal protein L31	AE000137.YKGM	0.392
BX842647.PE129	0.732	recA protein	AE000354.RECA	0.640
BX842654.RPSM	0.732	ribosomal protein S13p/S18e	AE000407.RPSM	0.548
BX842646.ATPB	0.723	ATP synthase F0, A subunit	AE000450.ATPB	0.556
BX842646.DPPA	0.720	ABC-type Dipeptide transport protein, periplasmic	AE000307.YEJA	0.409
BX842654.RPOA	0.715	DNA-directed RNA polymerase, alpha subunit	AE000407.RPOA	0.521
BX842647.PE282	0.713	glycine cleavage system H protein	AE000374.GCVH	0.542
BX842656.ICDA	0.694	isocitrate dehydrogenase, NADP-dependent	AE000213.ICDA	0.619
BX842652.CIT	0.693	GltA1	AE000140.PRPC	0.441
BX842653.YIAD	0.687	OmpA family protein	AE000432.YIAD	0.478
BX842651.PE151	0.682	butyryl-CoA dehydrogenase	AE000114.CAIA	0.489
BX842652.PE280	0.681	2-methylcitrate dehydratase	AE000140.PRPD	0.478
BX842646.CTAD	0.678	cytochrome c oxidase subunit I	AE000149.CYOB	0.543
BX842653.LEUC	0.677	aconitate hydratase, mitochondrial	AE000117.LEUC	0.490
BX842652.ALADH	0.676	alanine dehydrogenase	AE000255.PNTA	0.397
BX842656.PEPA	0.673	aminopeptidase A/I	AE000496.PEPA	0.482
BX842650.PE113	0.664	general secretion pathway protein G	AE000409.HOFG	0.389
BX842651.NRDA	0.652	ribonucleoside-diphosphate reductase alpha chain	AE000313.NRDA	0.574
BX842651.PE266	0.647	acetyl-CoA C-acetyltransferase	AE000311.ATOB	0.412
BX842656.ATPG	0.640	ATP synthase F1, gamma subunit	AE000450.ATPG	0.502
BX842654.SECY	0.636	preprotein translocase SecY subunit	AE000408.PRLA	0.507
BX842655.FLGE	0.636	flagellar hook protein FlgE	AE000208.FLGE	0.454
BX842650.HMOA	0.626	molybdopterin oxidoreductase, iron-sulfur binding subunit	AE000480.NRFC	0.376
BX842656.ATPD	0.620	ATP synthase F1, delta subunit	AE000450.ATPH	0.491
BX842654.NUOD	0.618	NADH dehydrogenase I,D subunit	AE000317.NUOC	0.519
BX842646.TALC	0.617	transaldolase, putative	AE000468.TALC	0.490
BX842651.PE338	0.616	butyryl-CoA dehydrogenase	AE000130.YAFH	0.433
BX842651.PE34	0.611	3-hydroxybutyryl-CoA dehydrogenase	AE000236.YDBU	0.395
BX842656.PCCB	0.607	propionyl-CoA carboxylase beta chain	AE000320.ACCD	0.527
BX842646.ETFA	0.606	Electron transfer flavoprotein alpha-subunit	AE000265.YDIR	0.427
BX842650.HIMA	0.606	integration host factor, alpha subunit	AE000266.HIMA	0.425
BX842647.PE314	0.605	glutamate dehydrogenase	AE000271.GDHA	0.495
BX842652.FBA	0.600	FBP aldolase, class I	AE000299.B2097	0.442
BX842653.TATA	0.600	twin-arginine-dependent translocase protein	AE000459.B3836	0.424
BX842649.DPPD	0.581	ABC-type dipeptide transport system, ATPase component	AE000185.B0829	0.391
BX842646.YDIY	0.579	conserved hypothetical protein	AE000267.B1722	0.420
BX842652.FOLE	0.576	GTP cyclohydrolase I (GTP-CH-I)	AE000304.FOLE	0.423

BX842648.PE202	0.572	S1 RNA binding domain protein	AE000416.YHGF	0.483
BX842656.PE37	0.572	spermidine synthase	AE000121.SPEE	0.455
BX842651.PE21	0.570	acetyl-CoA acyltransferase	AE000322.B2342	0.418
BX842647.PE290	0.566	glycine dehydrogenase	AE000373.GCVP	0.522
BX842646.DPPC	0.563	Dipeptide transport system permease protein	AE000307.YEJE	0.361
BX842650.LPXA	0.555	acyl-[acyl-carrier-protein]--UDP-N-acetylglucosamine O-acyltransferase	AE000127.LPXA	0.430
BX842655.UBIE	0.552	ubiquinone/menaquinone biosynthesis methyltransferase	AE000459.UBIE	0.437
BX842648.YCEI	0.551	Protein yceI precursor	AE000207.YCEI	0.435
BX842646.SUFC	0.550	FeS assembly ATPase SufC	AE000263.YNHD	0.345
BX842646.PE326	0.549	peptide ABC transporter, permease protein	AE000307.YEJB	0.379
BX842649.PUTA	0.549	1-pyrroline-5 carboxylate dehydrogenase	AE000203.PUTA	0.454
BX842652.PE281	0.549	2-methylisocitratelyase 2	AE000140.PRPB	0.412
BX842650.GSPD	0.547	general secretion pathway protein D	AE000409.YHEF	0.398
BX842656.PNP	0.547	purine nucleoside phosphorylase I, inosine and guanosine-specific	AE000328.XAPA	0.406
BX842646.BTUE	0.535	Vitamin B12 transport periplasmic protein btuE	AE000266.BTUE	0.396
BX842651.PE35	0.535	3-hydroxyacyl-CoA dehydrogenase and enoyl-CoA hydratase	AE000236.YDBS	0.397
BX842646.PMM	0.533	Phosphoglucomutase/phosphomannomutase, C-terminal domain family	AE000295.CPSG	0.416
BX842648.RPOE	0.529	RNA polymerase sigma-E factor	AE000343.RPOE	0.421
BX842649.AMN	0.529	AMP nucleosidase, putative	AE000290.AMN	0.367
BX842648.MAOC	0.528	maoC family protein	AE000236.MAOC	0.408
BX842648.LEPB	0.527	Signal peptidase I	AE000343.LEPB	0.444
BX842651.PE268	0.527	3-oxoacid CoA-transferase subunit B	AE000311.ATOA	0.340
BX842648.CMK	0.526	cytidylate kinase	AE000193.CMK	0.388
BX842652.SUGE	0.525	molecular chaperone sugE	AE000487.SUGE	0.369
BX842656.POLC	0.525	DNA polymerase III alpha subunit	AE000278.B1844	0.372
BX842649.MLTD	0.524	membrane-bound lytic murein transglycosylase D precursor	AE000130.DNIR	0.445
BX842655.RAGA	0.523	two component response regulator	AE000162.YLCA	0.337
BX842646.SODC	0.521	superoxide dismutase-like protein	AE000259.SODC	0.359
BX842652.PE228	0.517	phenol 2-monooxygenase	AE000459.UBIB	0.362
BX842656.CDSA	0.516	phosphatidate cytidylyltransferase	AE000127.CDSA	0.401
BX842646.GLNB	0.515	Nitrogen regulatory protein P-II	AE000341.GLNB	0.347
BX842655.MOTA	0.515	flagellar motor protein	AE000282.MOTA	0.411
BX842654.MURD	0.511	UDP-N-acetylmuramoylalanine--D-glutamate ligase	AE000118.MURD	0.435
BX842651.PE22	0.510	fatty oxidation complex, alpha subunit	AE000322.B2341	0.402
BX842650.PILR	0.507	regulator protein pilR	AE000311.ATOC	0.382
BX842648.PILQ	0.506	fimbrial assembly protein	AE000414.HOFQ	0.342
BX842650.PE118	0.505	predicted ATPases involved in pili biogenesis, PilB homologs	AE000409.YHEG	0.366
BX842651.RPIB	0.504	ribose 5-phosphate isomerase B	AE000482.RPIB	0.324
BX842656.METB	0.501	cystathionine gamma-lyase	AE000468.METB	0.433
BX842646.PE167	0.500	Transcriptional regulator superfamily	AE000490.YJEB	0.326
BX842656.HSDS	0.500	type I restriction-modification system, S subunit	AE000505.HSDS	0.394
BX842648.ETF-QO	0.498	electron transfer flavoprotein-ubiquinone oxidoreductase	AE000114.FIXC	0.433
BX842647.CHE	0.497	CinA-like protein	AE000354.YGAD	0.393
BX842651.PE267	0.495	3-oxoacid CoA-transferase subunit A	AE000311.ATOD	0.392
BX842652.PE151	0.494	THIF family protein	AE000364.YGDL	0.379
BX842655.CHEA	0.488	chemotaxis protein	AE000282.CHEA	0.410
BX842648.PHOH	0.486	PhoH-like ATPase	AE000204.PHOH	0.376
BX842646.PE38	0.483	radical activating enzyme	AE000361.YGCF	0.374

BX842655.GST	0.482	glutathione S-transferase family protein	AE000319.YFCF	0.387
BX842656.PILT	0.479	twitching motility protein	AE000378.YGGR	0.349
BX842648.TMK	0.478	thymidylate kinase	AE000210.TMK	0.376
BX842650.GIDB	0.478	Methyltransferase gidB	AE000451.GIDB	0.392
BX842652.PE248	0.475	soluble lytic murein transglycosylase	AE000379.MLTC	0.356
BX842655.PE273	0.473	phosphoesterase	AE000126.YAEI	0.387
BX842655.PE138	0.472	heme biosynthesis	AE000247.B1497	0.373
BX842649.PE87	0.465	glycine rich protein	AE000361.YGCG	0.367
BX842651.SUFE	0.465	Regulator of cysteine desulfurase activity	AE000263.YNHA	0.328
BX842650.PE21	0.463	oxidoreductase family protein	AE000207.MVIM	0.335
BX842647.FLIS	0.462	flagellar protein FliS	AE000285.FLIS	0.339
BX842648.MEPA	0.462	lipoprotein, putative	AE000321.MEPA	0.384
BX842650.PE117	0.462	general secretion pathway protein F	AE000409.HOFF	0.366
BX842651.PCAD	0.462	beta-ketoadipate enol-lactone hydrolase	AE000202.B1009	0.366
BX842646.ASTD	0.461	succinylglutamic semialdehyde dehydrogenase	AE000269.B1746	0.393
BX842654.PE236	0.445	conserved hypothetical protein	AE000238.B1410	0.309
BX842655.FLHA	0.444	flagellar biosynthesis protein FlhA	AE000281.FLHA	0.352
BX842648.PHOB	0.442	DNA-binding response regulator PhoB	AE000146.PHOB	0.345
BX842651.PE129	0.439	beta-lactamase	AE000330.B2430	0.364
BX842648.YGID	0.434	ortholog ygiD E.coli	AE000385.YGID	0.347
BX842647.TETA	0.433	multidrug resistance protein	AE000206.YCEE	0.370
BX842651.PE13	0.433	Acyl-CoA thioester hydrolase	AE000223.YCIA	0.313
BX842654.PE21	0.429	phosphatidate cytidylyltransferase	AE000238.B1409	0.293
BX842656.PE43	0.428	Aminopeptidase	AE000317.YFBL	0.351
BX842654.NADB	0.426	L-aspartate oxidase	AE000344.NADB	0.376
BX842646.PE3	0.424	DNA replication and repair protein RecF subfamily	AE000447.RECF	0.333
BX842655.BTUB	0.422	outer membrane receptor for transport of vitamin B12, E colicins, and bacteriophage BF23	AE000471.BTUB	0.362
BX842649.GLK	0.421	glucokinase	AE000145.YAJF	0.356
BX842650.PE263	0.420	quateRNAry ammonium compound-resistance protein qacE	AE000160.EMRE	0.286
BX842653.SSUC	0.417	ABC transporter , permease component	AE000195.YCBM	0.332
BX842652.APPC	0.411	ABC transporter, membrane spanning protein	AE000185.B0832	0.349
BX842649.PE309	0.410	transcriptional regulator, MarR family	AE000496.B4256	0.244
BX842653.PE81	0.397	cytochrome c oxidase accessory protein FixG	AE000201.YCCM	0.302
BX842650.PE304	0.396	Uncharacterized protein family UPF0061	AE000266.B1706	0.329
BX842654.GLPT	0.396	glycerol-3-phosphate transporter	AE000444.UHPC	0.332
BX842655.DSBD	0.395	thiol:disulfide interchange protein	AE000486.DSBD	0.345
BX842652.GLOB	0.392	hydroxyacylglutathione hydrolase GloB	AE000130.GLOB	0.311

Table of genes found to be potentially up-regulated in *E. coli* as compared to *B. bacteriovorus*.

Bdellovibrio Accession Number	*Bdellovibrio* Fop	Description	*E. coli* Accession Number	*E. coli* Fop
BX842646.SURA	0.444	PPIC-type PPIASE domain protein	AE000115.SURA	0.555
BX842646.PPIC	0.362	Parvulin-like peptidyl-prolyl isomerase	AE000150.YBAU	0.531
BX842646.LIG	0.378	DNA ligase, NAD-dependent	AE000328.LIG	0.472
BX842646.SERS	0.503	seryl-tRNA synthetase	AE000191.SERS	0.623
BX842646.AGLW	0.481	Adventurous gliding motility protein W	AE000177.TOLB	0.581
BX842646.SECA	0.535	preprotein translocase, SecA subunit	AE000119.SECA	0.609
BX842646.PHND	0.300	Phosphonates-binding protein	AE000482.PHND	0.425
BX842646.NDH	0.388	NADH dehydrogenase	AE000211.NDH	0.496
BX842646.PE331	0.505	leucyl-tRNA synthetase	AE000168.LEUS	0.596
BX842647.ACE	0.601	pyruvate dehydrogenase E1 component	AE000120.ACEE	0.699
BX842647.VACB	0.421	ribonuclease R	AE000490.VACB	0.495
BX842647.VALS	0.526	valyl-tRNA synthetase	AE000496.VALS	0.618
BX842647.LYSC	0.378	aspartate kinase	AE000475.LYSC	0.478
BX842647.PE179	0.335	inorganic pyrophosphatase	AE000494.PPA	0.704
BX842647.CHEA	0.325	histidine kinase	AE000282.CHEA	0.410
BX842647.PPC	0.338	conserved hypothetical protein	AE000469.PPC	0.494
BX842647.ASPC	0.383	aspartate aminotransferase	AE000318.B2290	0.504
BX842647.PE251	0.285	YjeF-related protein, C-terminus	AE000489.YJEF	0.390
BX842647.PE266	0.393	nucleoside-specific channel-forming protein tsx precursor	AE000147.TSX	0.678
BX842647.SLYD	0.414	FKBP-type peptidyl-prolyl cis-trans isomerase	AE000411.SLYD	0.683
BX842647.PE328	0.341	membrane protein, putative	AE000412.YHFC	0.438
BX842648.PGI	0.455	glucose-6-phosphate isomerase	AE000476.PGI	0.583
BX842648.PE22	0.317	methionine-R-sulfoxide reductase	AE000272.YEAA	0.479
BX842648.ENO	0.632	enolase	AE000361.ENO	0.826
BX842648.CKS	0.349	3-deoxy-D-manno-octulosonate cytidylyltransferase	AE000193.KDSB	0.489
BX842648.PYRG	0.422	CTP synthase	AE000361.PYRG	0.564
BX842648.GLNQ	0.358	similar to amino acid ABC transporter, ATP-binding protein	AE000183.GLNQ	0.544
BX842648.NRTD	0.347	ABC-type nitrate transporter, ATPase component	AE000399.YHBG	0.526
BX842648.APRT	0.348	adenine phosphoribosyltransferase	AE000153.APT	0.553
BX842648.METK	0.565	S-adenosylmethionine synthetase	AE000377.METK	0.645
BX842648.CARB	0.416	carbamoyl-phosphate synthase, large subunit	AE000113.CARB	0.555
BX842648.TOLC	0.351	outer membrane export factor	AE000385.TOLC	0.473
BX842648.HTRA	0.313	CBS domain protein	AE000125.HTRA	0.576
BX842648.PPK	0.341	polyphosphate kinase	AE000336.PPK	0.444
BX842648.CORA	0.382	magnesium and cobalt transport protein	AE000457.CORA	0.577
BX842648.GAPDH	0.634	glyceraldehyde-3-phosphate dehydrogenase, type I	AE000273.GAPA	0.800
BX842648.PGK	0.573	phosphoglycerate kinase	AE000376.PGK	0.751
BX842648.TPIA	0.505	triosephosphate isomerase	AE000466.TPIA	0.771
BX842648.ASNS	0.443	asparaginyl-tRNA synthetase	AE000195.ASNS	0.554
BX842648.NUSB	0.458	transcription antitermination factor NusB	AE000148.NUSB	0.600
BX842648.PE335	0.341	prolipoprotein diacylglyceryl transferase	AE000366.LGT	0.474
BX842649.PPDK	0.503	pyruvate, phosphate dikinase	AE000329.PTSI	0.609
BX842649.SPL1	0.450	Cysteine desulfurase (NifS protein homolog)	AE000339.YFHO	0.623
BX842649.NRDE	0.326	ribonucleoside-diphosphate reductase alpha chain	AE000352.NRDE	0.427

BX842649.RPSA	0.528	30S ribosomal protein S1	AE000193.RPSA	0.788
BX842649.CINA	0.324	nucleotide-utilizing enzyme	AE000315.B2249	0.465
BX842649.LAMB	0.444	maltoporin precursor	AE000477.LAMB	0.604
BX842649.LEUB	0.470	3-isopropylmalate dehydrogenase	AE000213.ICDA	0.619
BX842649.DNAK	0.620	Chaperone protein dnaK	AE000112.DNAK	0.737
BX842649.ACS	0.349	acetyl coenzyme A synthetase	AE000480.ACS	0.438
BX842649.PE197	0.507	ABC transporter, ATP-binding protein	AE000509.YJJK	0.599
BX842649.GLNS	0.474	glutaminyl-tRNA synthetase	AE000171.GLNS	0.585
BX842649.PEPP	0.385	aminopeptidase P	AE000374.PEPP	0.480
BX842649.PE255	0.404	TonB-dependent siderophore receptor, putative	AE000124.FHUA	0.512
BX842649.DEAD	0.564	ATP-dependent RNA helicase	AE000397.DEAD	0.650
BX842649.PE280	0.379	Spermidine/putrescine transport ATP-binding protein potA	AE000212.POTA	0.488
BX842649.POTB	0.343	spermidine/putrescine transport system permease protein	AE000212.POTB	0.465
BX842649.PE298	0.405	membrane protein	AE000305.YEIH	0.505
BX842649.PURA	0.495	adenylosuccinate synthetase	AE000490.PURA	0.656
BX842649.MVIN	0.312	integral membrane protein MviN	AE000208.MVIN	0.400
BX842650.PRFB	0.395	peptide chain release factor 2	AE000372.PRFB	0.550
BX842650.RIBC	0.320	riboflavin biosynthesis protein RibF	AE000113.RIBF	0.430
BX842650.UPP	0.397	uracil phosphoribosyltransferase	AE000336.UPP	0.576
BX842650.INFB	0.567	translation initiation factor IF-2	AE000397.INFB	0.644
BX842650.TRUB	0.344	tRNA pseudouridine synthase B	AE000397.TRUB	0.497
BX842650.MIAA	0.284	tRNA delta(2)-isopentenylpyrophosphate transferase	AE000489.MIAA	0.409
BX842650.PSTC	0.365	phosphate ABC transporter, permease protein PstC	AE000449.PSTC	0.500
BX842650.PSTB	0.485	phosphate ABC transporter, ATP-binding protein	AE000449.PSTB	0.599
BX842650.METG	0.480	methionyl-tRNA synthetase VC1036	AE000300.METG	0.571
BX842650.MREB	0.397	rod shape-determining protein	AE000404.MREB	0.600
BX842650.PE264	0.372	YieF	AE000448.YIEF	0.497
BX842650.DAPD	0.432	2,3,4,5-tetrahydropyridine-2,6-dicarboxylate N-succinyltransferase(EC 2.3.1.117) (Tetrahydrodipicolinate N-succinyltransferase)(THP succinyltransferase) (Tetrahydropicolinate succinylase)	AE000126.DAPD	0.540
BX842650.GUAB	0.452	Inosine-5-monophosphate dehydrogenase	AE000337.GUAB	0.652
BX842650.HTPG	0.488	heat shock protein htpG	AE000153.HTPG	0.632
BX842651.PSD	0.395	phosphatidylserine decarboxylase	AE000488.PSD	0.522
BX842651.FEOB	0.384	ferrous iron transport protein B	AE000416.FEOB	0.502
BX842651.PE59	0.308	ABC-type multidrug transporter permease protein	AE000181.YBHS	0.419
BX842651.PE68	0.363	2-oxoglutarate/malate translocator	AE000179.YBHI	0.478
BX842651.FKPA	0.327	peptidyl-prolyl cis-trans isomerase, FKBP-type	AE000410.FKPA	0.640
BX842651.FTSH	0.516	cell division protein	AE000398.HFLB	0.607
BX842651.PURB	0.442	adenylosuccinate lyase	AE000213.PURB	0.570
BX842651.FABG	0.403	oxidoreductase	AE000210.FABG	0.536
BX842651.GLYA	0.503	serine hydroxymethyltransferase	AE000341.GLYA	0.659
BX842651.ARGS	0.462	arginyl-tRNA synthetase	AE000281.ARGS	0.570
BX842651.DACA	0.321	InterPro: D-alanyl-D-alanine carboxypeptidase 1	AE000186.DACC	0.431
BX842651.RBSB	0.388	D-ribose periplasmic binding protein	AE000452.RBSB	0.521
BX842651.GUAA	0.446	GMP synthase	AE000337.GUAA	0.614
BX842651.GUAB	0.415	inosine-5-monophosphate dehydrogenase	AE000337.GUAB	0.652
BX842651.PE259	0.378	tRNA-i(6)A37 thiotransferase enzyme MiaB	AE000170.YLEA	0.483
BX842651.PYKA	0.408	pyruvate kinase	AE000279.PYKA	0.543

BX842651.HUA	0.415	DNA-binding protein HU-alpha	AE000473.HUPA	0.705
BX842651.SRP54	0.389	signal recognition particle protein	AE000347.FFH	0.507
BX842652.COPA	0.342	copper-translocating P-type ATPase	AE000154.YBAR	0.438
BX842652.GLTX	0.501	glutamyl-tRNA synthetase	AE000328.GLTX	0.581
BX842652.PE128	0.423	protease	AE000298.YEGQ	0.512
BX842652.FUR	0.413	ferric uptake regulation protein	AE000172.FUR	0.549
BX842652.LPXC	0.407	UDP-3-0-acyl N-acetylglucosamine deacetylase	AE000119.LPXC	0.518
BX842652.AHPF	0.472	alkyl hydroperoxide reductase, subunit F	AE000166.AHPF	0.568
BX842653.PE60	0.443	transcription regulator containing cAMP-binding domain	AE000411.CRP	0.568
BX842653.PE70	0.382	transcriptional regulator Crp/FNR family	AE000231.FNR	0.541
BX842653.FECA	0.348	outer membrane iron	AE000499.FECA	0.458
BX842653.PFS	0.408	MTA/SAH nucleosidase	AE000125.PFS	0.533
BX842653.PYRC	0.437	dihydroorotase	AE000157.YBBX	0.603
BX842653.PE162	0.370	merozoite surface protein-3a	AE000177.TOLA	0.529
BX842653.PE168	0.322	GTPase	AE000488.YJEQ	0.471
BX842653.PRS	0.397	ribose-phosphate pyrophosphokinase	AE000219.PRSA	0.639
BX842653.GLPK	0.405	glycerol kinase	AE000467.GLPK	0.494
BX842653.GPMA	0.498	phosphoglycerate mutase	AE000178.GPMA	0.624
BX842653.LPDA	0.570	dihydrolipoamide dehydrogenase	AE000121.LPDA	0.688
BX842653.PE219	0.358	Predicted permease YjgP/YjgQ family superfamily	AE000497.YJGQ	0.458
BX842653.HTPX	0.400	protease heat shock protein	AE000277.HTPX	0.538
BX842653.PE273	0.336	mechanosensitive ion channel family protein	AE000152.AEFA	0.483
BX842653.NRFA	0.378	nitrite reductase periplasmic cytochrome c552	AE000480.NRFA	0.506
BX842653.PE296	0.274	Uncharacterized protein conserved in bacteria	AE000179.YBHH	0.394
BX842653.PE308	0.328	Uvs121	AE000418.RTCB	0.424
BX842653.ASPA	0.428	aspartate ammonia-lyase	AE000486.ASPA	0.700
BX842654.SERA	0.408	D-3-phosphoglycerate dehydrogenase	AE000374.SERA	0.512
BX842654.YCBB	0.299	putative amidase	AE000194.YCBB	0.420
BX842654.MUTS	0.332	MutS-like mismatch repair protein, ATPases	AE000357.MUTS	0.452
BX842654.ADK	0.456	adenylate kinase	AE000153.ADK	0.656
BX842654.RPSC	0.648	ribosomal protein S3	AE000408.RPSC	0.761
BX842654.FUSA	0.493	translation elongation factor G	AE000410.FUSA	0.787
BX842654.RPOC	0.674	DNA-directed RNA polymerase, beta subunit	AE000472.RPOC	0.729
BX842654.RPLK	0.639	ribosomal protein L11	AE000472.RPLK	0.775
BX842654.PURM	0.368	phosphoribosylformylglycinamidine cyclo-ligase	AE000336.PURM	0.474
BX842654.PURH	0.390	IMP cyclohydrolase	AE000473.PURH	0.500
BX842654.PURL	0.398	phosphoribosylformylglycinamidine synthase II	AE000342.PURL	0.526
BX842654.PURD	0.344	phosphoribosylamine--glycine ligase	AE000473.PURD	0.455
BX842654.PURF	0.304	amidophosphoribosyltransferase	AE000320.PURF	0.500
BX842654.PROS	0.480	prolyl tRNA synthetase	AE000128.PROS	0.616
BX842654.CLPB	0.384	ATPase with chaperone activity, two ATP-binding domains	AE000345.CLPB	0.527
BX842654.XERD	0.310	site-specific recombinase	AE000457.XERC	0.420
BX842654.FABD	0.290	malonyl CoA-acyl carrier protein transacylase	AE000210.FABD	0.516
BX842654.PPSA	0.378	phosphoenolpyruvate synthase	AE000265.PPSA	0.534
BX842654.RECN	0.384	DNA repair protein RecN	AE000347.RECN	0.470
BX842654.MURC	0.384	UDP-N-acetylmuramate--alanine ligase	AE000118.MURC	0.469
BX842655.MRAW	0.359	S-adenosyl-methyltransferase MraW	AE000118.YABC	0.463
BX842655.PE70	0.299	putative polysaccharide deacetylase	AE000135.YAGG	0.424
BX842655.HSP	0.351	molecular chaperone, Hsp70 family	AE000297.YEGD	0.487
BX842655.ASPS	0.508	aspartyl-tRNA synthetase	AE000280.ASPS	0.652

BX842655.RP32	0.348	RNA polymerase sigma-32 factor	AE000422.RPOH	0.614
BX842655.SUHB	0.379	inositol-1-monophosphatase	AE000339.SUHB	0.618
BX842655.PCRA	0.385	ATP-dependent DNA helicase	AE000457.UVRD	0.466
BX842655.PE160	0.482	Phosphofructokinase	AE000466.PFKA	0.693
BX842655.FLGC	0.295	flagellar basal-body rod protein FlgC	AE000208.FLGC	0.455
BX842655.GLMS	0.436	glucosamine--fructose-6-phosphate aminotransferase, isomerizing	AE000450.GLMS	0.586
BX842655.SPEB	0.408	agmatinase, putative	AE000377.SPEB	0.525
BX842655.PE229	0.275	LrgA family holin protein	AE000303.YOHJ	0.432
BX842655.PE233	0.328	putative hydrolase	AE000172.YBFF	0.494
BX842655.PE245	0.490	Transketolase, pyridine binding domain protein	AE000376.TKTA	0.705
BX842655.MENF	0.327	phospho-2-dehydro-3-deoxyheptonate aldolase	AE000316.MENF	0.432
BX842655.MENB	0.446	naphthoate synthase	AE000316.MENB	0.624
BX842655.PANB	0.366	3-methyl-2-oxobutanoate hydroxymethyltransferase	AE000122.PANB	0.521
BX842656.PE58	0.434	ABC transporter ATP-binding protein	AE000184.YBIT	0.531
BX842656.FDX	0.266	2Fe-2S ferredoxin	AE000339.FDX	0.571
BX842656.DRA	0.439	deoxyribose-phosphate aldolase	AE000508.DEOC	0.598
BX842656.DNAJ	0.406	DnaJ protein	AE000112.DNAJ	0.529
BX842656.YIDC	0.519	60 KD inner-membrane protein	AE000447.YIDC	0.614

Predicting Gene Expression Level from Codon Usage Bias

Ian Henry and Paul M. Sharp

Institute of Genetics, University of Nottingham, Queens Medical Centre, Nottingham, United Kingdom

The "expression measure" of a gene, E(g), is a statistic devised to predict the level of gene expression from codon usage bias. E(g) has been used extensively to analyze prokaryotic genome sequences. We discuss 2 problems with this approach. First, the formulation of E(g) is such that genes with the strongest selected codon usage bias are not likely to have the highest predicted expression levels; indeed the correlation between E(g) and expression level is weak among moderate to highly expressed genes. Second, in some species, highly expressed genes do not have unusual codon usage, and so codon usage cannot be used to predict expression levels. We outline a simple approach, first to check whether a genome shows evidence of selected codon usage bias and then to assess the strength of bias in genes as a guide to their likely expression level; we illustrate this with an analysis of *Shewanella oneidensis*.

When *Escherichia coli* gene sequence data began to accumulate, it became apparent that alternative synonymous codons are not used with equal frequencies. Translationally optimal codons can be identified as those best recognized by the most abundant tRNAs, and the frequency of these codons in a gene is highly correlated with gene expression level (Post and Nomura 1980; Ikemura 1981; Gouy and Gautier 1982). It follows that the strength of codon usage bias in a gene can be used to make predictions about its expression level. Karlin and colleagues have devised a codon usage statistic termed E(g), the "expression measure" of a gene, which they have used in attempts to identify predicted highly expressed (PHX) genes in a wide range of prokaryotic genomes (Karlin and Mrazek 2000, 2001; Karlin et al. 2001, 2003, 2004, 2005, 2006; Mrazek et al. 2001, 2006; Ma et al. 2002). Here we discuss problems with the E(g) statistic and with its application to diverse species.

To verify the utility of their approach, Karlin et al. (2001) compared E(g) values with protein abundance data from 2D gel electrophoresis for 96 *E. coli* genes. The protein relative molecular abundance (RMB) values varied from 0.116 to 41.8, whereas the E(g) values ranged from 0.38 to 2.66. Analysis of the values given by Karlin et al. (2001) shows that although there is an overall correlation between E(g) and (the log of) protein abundance, it is quite weak (0.41). Among the 18 proteins with the lowest RMB values were 7 encoded by genes with E(g) values greater than 1.0, the criterion used to classify a gene as PHX. Furthermore, the highest E(g) value was that for the *pnp* gene, encoding polynucleotide phosphorylase which, with an RMB of 1.22, ranked only 48 out of 96 for protein abundance. This E(g) value for *pnp* was also the highest among all genes in the *E. coli* K-12 genome (Karlin and Mrazek 2000).

Two previously described statistics, the frequency of optimal codons (F_{OP}; Ikemura 1985) and the codon adaptation index (CAI; Sharp and Li 1987), do not give such strange results. For example, among these 96 genes, the gene with the highest CAI value (0.84) is *rplL* encoding ribosomal protein L7/12, one of *the* most abundant proteins in *E. coli*, especially under the rapid growth conditions when codon selection is expected to be effective. The highest CAI value among all *E. coli* genes is 0.85 for *lpp*, encoding an outer membrane lipoprotein that is *the* most abundant protein in the *E. coli* cell (DiRienzo et al. 1978). The CAI for *pnp* is 0.63, indicating above average, but not extreme, selected codon usage bias. Overall, the correlation of CAI and log(RMB) among the 96 genes is 0.53.

Consideration of the nature of these various measures of codon usage bias can explain the differences in these results. F_{OP} is simply the frequency of the optimal codons within a gene (Ikemura 1985); the CAI is similar, but weights suboptimal codons differentially, according to the extent of their avoidance in very highly expressed genes (Sharp and Li 1987). With either approach, the value increases with greater bias to more optimal codons, up to a potential maximum of 1.0 when only the best codon for each amino acid is used. In contrast, it can be seen that genes with the most strongly selected codon usage bias, reflecting the highest gene expression, are not expected to have the highest E(g) values. To calculate E(g) for gene X, Karlin takes the codon usage of that gene (X), of the genome as a whole (G), and of a reference set of genes expected to be expressed at high levels (H). The equation for E(g) takes the general form d_{XG}/d_{XH}, where the terms are the (absolute) differences in codon usage between gene X and either the genome as a whole (d_{XG}) or the highly expressed genes (d_{XH}). Thus E(g) will be higher when d_{XH} is smaller, and the maximum E(g) value would be achieved when the codon usage of gene X matches the overall codon usage of the reference set H. Since this reference set contains numerous genes, codon usage summed across it does not have the strongest possible bias. Consequently, d_{XH} is at a minimum in genes with less than extreme codon usage bias and increases (making E(g) lower) in genes with stronger selected codon usage bias than the reference set H.

Although genes expressed at low levels should have low E(g) values, the values for genes expressed at moderate to high levels are less predictable. Among the 96 genes discussed above, for the 48 encoding proteins with above median RMB values the correlation of E(g) with log(RMB) is only 0.22; in contrast, the correlation of CAI with log(RMB) is 0.48. The difference is due to a large number of genes with anomalously high E(g) values given their moderate expression levels (fig. 1). For both measures of codon usage bias, the correlation is weakened by the values for *metE*. The metE protein was very abundant under the growth conditions used to obtain the RMB values but is expressed at a 50-fold lower level under rapid growth conditions (Pedersen et al. 1978). When *metE* is excluded, the correlation of E(g) and log(RMB) is still only 0.30, whereas that for CAI and log(RMB) is 0.58.

Key words: codon usage, gene expression, predicted highly expressed genes, *Escherichia coli*, *Shewanella oneidensis*.

E-mail: paul@evol.nott.ac.uk.

Mol. Biol. Evol. 24(1):10–12. 2007
doi:10.1093/molbev/msl148
Advance Access publication October 12, 2006

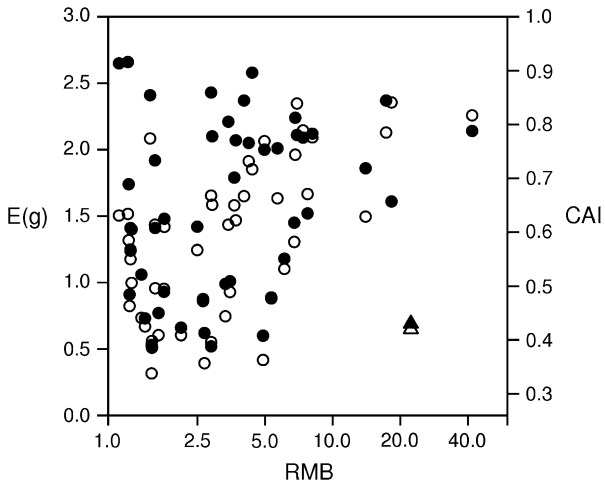

FIG. 1.—Codon bias measures for *Escherichia coli* genes encoding proteins with above average abundance (RMB). Black points: *E*(g), the expression measure (left axis). White points: CAI, the codon adaptation index (right axis). Triangles denote values for *metE* (see text).

As well as noting individual gene *E*(g) values, Karlin and colleagues have used *E*(g) to identify PHX genes. In one sense this is less problematic because, even for very strongly biased genes, *E*(g) values are unlikely to decrease below 1.0. However, this categorization of genes brings a different problem because using an arbitrary threshold value of *E*(g) must lead to genes with very similar codon usage bias, but lying on either side of this threshold, being classified as PHX and non-PHX, respectively.

A further problem arises when the *E*(g)/PHX methodology is applied to other species. The strength of selected codon usage bias varies widely among bacteria. In some species, such as *Helicobacter pylori* (Lafay et al. 2000); the mollicutes, *Mycoplasma genitalium* and *M. pneumoniae* (Kerr et al. 1997); or the spirochetes, *Borrelia burgdorferi* and *Treponema pallidum* (Lafay et al. 1999), highly expressed genes have no discernible difference in codon usage from other genes. In a recent survey of 80 bacterial genomes, we found 30% to show no significant evidence of translationally selected codon usage bias (Sharp et al. 2005). Clearly, in the absence of selected codon usage bias, the expression levels of genes are unlikely to be predictable from comparisons of codon usage. However, Karlin and colleagues have used their approach to study numerous species with little or no evidence of selected codon usage bias, including those listed above as well as *Rickettsia prowazekii*, *Chlamydia trachomatis*, *Chlamydophila pneumoniae*, *Blochmannia floridanus*, and *Buchnera* species (Karlin and Mrazek 2000; Mrazek et al. 2006). In such species, *E*(g) values may still vary among genes, reflecting stochastic variation in codon usage or systematic effects unrelated to gene expression level. For example, *B. burgdorferi* and *T. pallidum* exhibit a very strong base composition skew between the leading and lagging strands of replication (Lafay et al. 1999). Because highly expressed genes generally lie on the leading strand, other genes on this strand are likely to be given higher *E*(g) values, whether or not they are highly expressed.

In conclusion, to estimate the level of gene expression from codon usage bias, it is necessary first to establish whether highly expressed genes have translationally selected biased codon usage, and then (if they do) it seems most appropriate to apply a statistic that is maximized when that selected bias is maximized. The first step is easily achieved by comparing the codon usage of a standard set of highly expressed genes with that in the genome as a whole. It is then a simple matter to calculate the frequency of optimal codons in each gene.

As an example, we have analyzed *Shewanella oneidensis*, a member of the gamma proteobacteria (Heidelberg et al. 2002). Eighteen codons, for 15 amino acids, occur at significantly higher frequencies in highly expressed genes than in the genome as a whole (see Supplementary Material online). Importantly, these codons do not reflect any simple compositional bias, such as G + U richness due to location of the highly expressed genes on the leading strand of replication. Rather, they include many codons which would be expected to be optimal, either because they are decoded by the most abundant tRNA species (e.g., 6 of the 9 Arg tRNA genes match CGU) or because they are perfectly complementary to the only tRNA species for the amino acid (e.g., UUC, UAC, CAC, AAC, GAC, and GAA). F_{OP} values can be calculated for each gene as the frequency of these 18 codons among all codons for these 15 amino acids. F_{OP} values range from 0.09 in SO0711, a short (32 codon) hypothetical gene, to 0.89 in SO2787, encoding a cold shock protein. The top 20 scoring genes, with $F_{OP} > 0.72$, include 13 encoding ribosomal proteins and 4 encoding translation elongation factors. In contrast, Mrazek et al. (2006) found the highest *E*(g) values in *acnB* and *rpoB*, which rank outside the top 70 genes, with F_{OP} values of 0.60 and 0.62, respectively. Among the 185 PHX genes identified by Mrazek et al. (2006), the minimum F_{OP} value is 0.46; we find another 139 genes, each at least 80 codons long, with higher frequencies of optimal codons, that were not identified as PHX.

Finally, there are limitations to the use of codon usage bias in estimating gene expression levels. For example, if the selection pressures are stronger on genes expressed at higher levels during rapid growth, genes highly expressed only under other growth conditions cannot be detected—the *metE* gene of *E. coli* discussed above seems to be such a case.

Supplementary Material

A supplementary table is available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgment

I.H. was supported by a Biotechnology and Biological Sciences Research Council studentship.

Literature Cited

DiRienzo JM, Nakamura K, Inouye M. 1978. The outer membrane proteins of gram-negative bacteria: biosynthesis, assembly, and functions. Annu Rev Biochem. 47:481–532.

Gouy M, Gautier C. 1982. Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res. 10:7055–7074.

Heidelberg J, Paulsen I, Nelson K, et al. (43 co-authors). 2002. Genome sequence of the dissimilatory metal ion-reducing bacterium *Shewanella oneidensis*. Nat Biotechnol. 20:1118–1123.

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

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

Karlin S, Barnett MJ, Campbell A, Fisher RF, Mrazek J. 2003. Predicting gene expression levels from codon biases in α-proteobacterial genomes. Proc Natl Acad Sci USA. 100: 7313–7318.

Karlin S, Brocchieri L, Mrazek J, Kaiser D. 2006. Distinguishing features of δ-proteobacterial genomes. Proc Natl Acad Sci USA. 103:11352–11357.

Karlin S, Mrazek J. 2000. Predicted highly expressed genes of diverse prokaryotic genomes. J Bacteriol. 182:5238–5250.

Karlin S, Mrazek J. 2001. Predicted highly expressed and putative alien genes of *Deinococcus radiodurans* and implications for resistance to ionizing radiation damage. Proc Natl Acad Sci USA. 98:5240–5245.

Karlin S, Mrazek J, Campbell A, Kaiser D. 2001. Characterizations of highly expressed genes of four fast-growing bacteria. J Bacteriol. 183:5025–5040.

Karlin S, Mrazek J, Ma J, Brocchieri L. 2005. Predicted highly expressed genes in archaeal genomes. Proc Natl Acad Sci USA. 102:7303–7308.

Karlin S, Theriot J, Mrazek J. 2004. Comparative analysis of gene expression among low G+C gram-positive genomes. Proc Natl Acad Sci USA. 101:6182–6187.

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

Lafay B, Atherton JC, Sharp PM. 2000. Absence of translationally selected codon usage bias in *Helicobacter pylori*. Microbiology. 146:851–860.

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. 27:1642–1649.

Ma J, Campbell A, Karlin S. 2002. Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures. J Bacteriol. 184:5733–5745.

Mrazek J, Bhaya D, Grossman AR, Karlin S. 2001. Highly expressed and alien genes of the *Synechocystis* genome. Nucleic Acids Res. 29:1590–1601.

Mrazek J, Spoorman AM, Karlin S. 2006. Genomic comparisons among γ-proteobacteria. Environ Microbiol. 8:273–288.

Pedersen S, Bloch PL, Reeh S, Neidhardt FC. 1978. Patterns of protein synthesis in *E. coli*: a catalog of the amount of 140 individual proteins at different growth rates. Cell. 14:179–190.

Post LE, Nomura M. 1980. DNA sequences from the str operon of *Escherichia coli*. J Biol Chem. 255:4660–4666.

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

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

William Martin, Associate Editor

Accepted October 10, 2006

		Strength of Selected Codon Usage Bias for Each Amino Acid																							
Bacterium	Group	WWY	Phe	Tyr	Asn	Ile	Glu	His	Cys	Gln	Asp	Lys	Val AU/GC	Val AG/UC	Pro AU/GC	Pro AG/UC	Thr AU/GC	Thr AG/UC	Ala AU/GC	Ala AG/UC	Gly AU/GC	Gly AG/UC	Accession Number	Species
Agrtum	Alpha Pr	1.048	0.193	0.246	0.331	0.277	0.686	-1.462	-2.025	-2.402	-1.210	-1.846	1.368	-1.017	-0.154	0.557	-0.159	-0.404	1.489	-0.481	0.257	-2.258	AE008688*	Agrobacterium tumefaciens C58 (UW)
Anamar	Alpha Pr	-0.083	0.008	-0.011	-0.016	-0.064	-0.522	-0.166	0.457	-0.764	0.336	-0.736	-0.023	-0.196	-0.195	-0.523	-0.099	-0.475	-0.135	-0.563	0.377	-0.315	CP000030	Anaplasma marginale
Barhen	Alpha Pr	-0.373	-0.028	-0.059	-0.106	-0.180	-0.495	0.085	0.991	-0.830	0.389	-0.602	0.209	-0.322	-0.095	0.226	-0.235	-0.110	0.090	-0.148	0.369	-0.670	BX897699	Bartonella henselae
Barqui	Alpha Pr	-0.315	-0.055	-0.047	-0.055	-0.158	-0.580	-0.074	0.982	-0.734	0.492	-0.619	0.240	-0.347	-0.063	0.267	-0.262	-0.105	0.134	-0.098	0.338	-0.686	BX897700	Bartonella quintana
Brajap	Alpha Pr	0.741	0.284	0.110	0.219	0.129	0.256	-1.038	-0.247	-0.893	-0.465	-1.417	-0.104	-0.234	-1.132	0.415	-0.814	-0.626	-0.184	-0.345	0.198	-1.428	BA000040	Bradyrhizobium japonicum
Brumel	Alpha Pr	0.896	0.200	0.210	0.274	0.211	0.657	-1.084	-2.054	-2.106	-0.538	-1.741	0.761	-0.893	-0.262	0.645	-0.650	-0.458	0.904	-0.708	0.588	-1.835	AE008917*	Brucella melitensis
Caucre	Alpha Pr	1.152	0.298	0.207	0.319	0.328	1.408	-1.871	-0.880	1.254	-0.656	-0.661	0.558	-0.495	-1.158	0.691	-0.572	-0.290	1.049	-0.993	0.520	-2.799	AE005673	Caulobacter crescentus
Ehrcan	Alpha Pr	-0.765	-0.108	-0.170	-0.310	-0.177	-0.526	0.679	0.102	-0.352	0.904	-0.656	0.182	-0.215	-0.327	-0.428	-0.013	-0.556	0.179	-0.281	-0.243	-0.339	CP000107	Ehrlichia canis Jake
Ehrrum	Alpha Pr	-0.673	-0.169	-0.171	-0.178	-0.154	-0.679	0.636	0.797	-0.580	0.512	-0.760	-0.088	-0.208	-0.357	-0.396	0.307	-0.446	-0.254	-0.413	-0.306	-0.387	CR767821	Ehrlichia ruminatium strain Welgevonden
Gluoxy	Alpha Pr	0.785	0.207	0.132	0.322	0.123	-0.180	-1.118	-2.583	-1.273	-0.575	-2.197	0.609	-0.385	-0.461	0.851	-1.054	-0.068	0.957	-0.614	0.424	-1.878	CP000009	Gluconobacter oxydans
Meslot	Alpha Pr	0.757	0.290	0.132	0.166	0.169	0.155	-0.929	-0.761	-2.090	-0.425	-1.499	0.115	-0.549	-1.053	0.677	-0.887	-0.263	0.262	-0.449	0.098	-1.604	BA000012	Mesorhizobium loti
Pelubi	Alpha Pr	0.353	0.077	0.105	0.117	0.054	0.322	-0.850	-0.168	0.024	-0.104	-0.020	0.548	-0.205	0.541	-0.025	0.366	0.095	0.626	-0.148	0.226	-0.379	CP000084	Pelagibacter ubique
Rhopal	Alpha Pr	0.505	0.239	0.096	0.226	-0.056	-0.077	-1.170	-0.998	-1.876	-0.535	-1.827	0.142	-0.340	-1.300	0.771	-0.313	-0.593	0.015	-0.280	0.412	-1.508	BX571963	Rhodopseudomonas palustris
Riccon	Alpha Pr	-0.410	-0.059	-0.044	-0.210	-0.097	0.379	-0.408	0.421	0.089	0.446	0.216	0.235	-0.011	0.248	-0.349	0.183	-0.079	0.295	-0.019	0.485	-0.085	AE006914	Rickettsia conorii
Ricpro	Alpha Pr	-0.421	-0.115	-0.059	-0.139	-0.108	0.385	0.277	0.664	-0.234	0.507	0.013	0.018	-0.135	0.294	-0.486	0.085	-0.254	0.205	-0.116	0.172	-0.091	AJ235269	Rickettsia prowazekii
Rictyp	Alpha Pr	-0.460	-0.082	-0.043	-0.199	-0.136	0.058	-0.108	0.752	-0.242	0.603	0.045	0.084	-0.215	0.126	-0.408	0.139	-0.228	0.213	-0.123	0.150	-0.214	AE017197	Rickettsia typhi
Silpom	Alpha Pr	0.925	0.275	0.137	0.373	0.140	0.777	-1.637	-1.484	-0.941	-0.718	0.045	0.457	-0.650	-0.521	0.066	-0.501	-1.136	1.072	-0.509	0.576	-2.038	CP000031	Silicibacter pomeroyi
Sinmel	Alpha Pr	0.637	0.208	0.137	0.212	0.080	0.788	-1.210	-1.776	-2.086	-0.419	-1.551	0.815	-0.620	-0.917	0.852	-0.584	-0.181	0.816	-0.422	0.240	-2.123	AL591688	Sinorhizobium Meliloti
Wolpip	Alpha Pr	-0.684	-0.220	-0.113	-0.174	-0.177	-0.283	0.702	0.680	0.013	0.753	-0.466	0.023	-0.101	0.746	-0.899	0.137	-0.245	-0.045	-0.308	0.179	-0.495	AE017196	Wolbachia pipientis
Woltrs	Alpha Pr	-0.574	-0.177	-0.061	-0.211	-0.125	-0.249	0.537	0.865	0.112	0.513	-0.363	0.125	0.041	-0.051	-0.726	0.218	-0.008	0.119	-0.232	0.277	-0.287	AE017321	Wolbachia strain TRS
Zymmob	Alpha Pr	0.750	0.206	0.077	0.254	0.214	0.461	-0.596	-0.722	-2.151	-0.397	-0.536	0.327	-0.211	-0.573	0.653	-0.647	-0.368	0.843	-0.463	0.659	-1.907	AE008692	Zymomonas mobilis ZM4
Azoebn	Beta Pr	-0.055	0.020	-0.027	-0.018	-0.031	0.325	-0.195	-0.402	-0.111	0.189	-0.492	0.234	-0.115	-0.835	0.354	-0.345	-0.351	0.276	-0.295	0.500	-0.669	CR555306	Azoarcus sp. EbN1
Borper	Beta Pr	-0.033	0.042	0.064	0.124	-0.263	0.982	-0.906	-2.666	0.877	0.190	-0.102	1.054	-0.649	-0.272	-0.128	0.082	-0.219	1.397	-0.747	1.276	-1.504	BX470248	Bordertella pertussis
Burpse	Beta Pr	0.340	0.041	0.111	0.130	0.059	0.972	-1.081	-2.170	0.366	-0.339	-0.774	0.815	-0.293	-0.349	0.491	-0.221	-0.085	1.391	-0.545	1.102	-1.539	BX571965*	Burkholderia pseudomallei
Chrvio	Beta Pr	0.545	0.109	0.141	0.254	0.041	1.111	-0.941	-0.631	0.937	0.171	-0.190	1.883	-0.491	-0.232	1.100	0.798	-1.591	1.950	-0.934	1.390	-2.426	AL646052*	Chromobacterium violaceum
Decaro	Beta Pr	0.323	0.165	0.056	0.104	-0.002	0.500	-0.506	0.258	-0.149	-0.012	-1.094	0.961	-0.398	-0.621	0.599	-0.169	-0.843	1.002	-0.374	0.938	-1.744	CP000089	Dechloromonas aromatica
Neimen	Beta Pr	-0.099	0.095	0.091	-0.080	-0.204	-0.267	0.256	0.805	1.407	0.404	0.451	1.351	0.156	1.424	-0.329	1.005	-1.204	1.352	-0.339	1.408	-1.809	AL157959	Neisseria meningitidis Z2491
Niteur	Beta Pr	-0.884	-0.205	-0.149	-0.170	-0.360	-0.324	0.585	0.774	0.310	0.471	-0.003	0.791	-0.002	0.949	-0.060	0.571	0.356	0.765	0.096	0.555	0.233	AL954747	Nitrosomonas europaea
Raleut	Beta Pr	0.675	0.197	0.214	0.219	0.046	0.662	-1.162	-1.322	0.535	-0.561	-0.739	0.474	-0.379	-0.756	0.597	-0.671	-0.920	0.960	-0.957	0.847	-2.397	CP000090*	Ralstonia eutropha JMP134
Ralsol	Beta Pr	0.024	0.091	0.054	0.024	-0.145	1.111	-0.744	-1.012	1.027	-0.454	-1.097	0.345	-0.242	-0.726	0.747	-0.205	-0.392	1.254	-0.523	1.182	-2.253	AL646052*	Ralstonia solanacearum
Aciadp	Gamma Pr	1.545	0.359	0.228	0.482	0.477	0.374	-1.527	-0.551	1.127	-1.184	0.152	1.368	-0.339	0.753	0.259	1.149	-0.596	0.727	-0.449	0.829	-2.611	CR543861	Acinetobacter sp.ADP1
Bloflo	Gamma Pr	-1.067	-0.183	-0.279	-0.280	-0.326	-0.583	1.477	0.785	-0.819	0.783	-0.686	-0.031	-0.170	0.062	-0.201	0.088	-0.461	-0.224	-0.210	-0.241	-0.166	BX248583	Blochmannia floridandus
Blopen	Gamma Pr	-0.074	-0.010	-0.046	-0.021	0.003	-0.269	0.157	-0.208	-0.111	0.059	-0.250	0.055	0.058	0.016	-0.110	0.048	-0.343	0.034	-0.053	-0.240	-0.201	CP000016	Blochmannia pennsylvanicus
Buchap	Gamma Pr	-0.017	-0.018	-0.034	0.048	-0.013	-0.395	0.069	0.422	-0.270	0.192	-0.280	-0.200	-0.022	0.166	-0.052	0.158	-0.360	-0.085	-0.086	-0.004	-0.529	BA000003	Buchnera aphidicola Ap
Buchbp	Gamma Pr	-0.590	-0.072	-0.134	-0.128	-0.257	-0.422	1.181	0.431	-0.410	0.719	-0.416	0.010	-0.081	-0.273	-0.180	0.031	-0.310	-0.163	-0.121	-0.065	-0.451	AF492592	Buchnera aphidicola Bp
Buchsg	Gamma Pr	-0.069	-0.061	-0.018	-0.016	0.025	-0.143	-0.406	-0.128	-0.329	-0.016	-0.374	0.301	-0.150	0.305	0.041	-0.121	-0.395	-0.110	-0.187	-0.178	-0.349	AE013218	Buchnera aphidicola Sg
Colpsy	Gamma Pr	1.344	0.385	0.175	0.326	0.458	0.685	-0.993	1.868	0.530	-0.404	-0.092	1.431	-0.334	1.638	0.191	1.482	-0.250	0.904	-0.131	0.601	-1.542	CP000083	Colwellia psychrerythraea
Coxbur	Gamma Pr	0.175	-0.003	0.026	0.097	0.055	-0.139	-0.428	0.371	-0.300	-0.388	-0.299	-0.265	0.433	-0.182	0.450	-0.316	0.287	-0.010	0.207	-0.132	-0.302	AE016828	Coxiella burnetti
Erwcar	Gamma Pr	0.951	0.235	0.144	0.259	0.313	0.297	-0.580	0.051	0.224	-0.499	0.024	1.611	-0.413	0.363	0.776	1.014	-1.697	1.235	-0.431	0.956	-2.459	BX950851	Erwinia carotovora
Esccol	Gamma Pr	1.489	0.321	0.202	0.416	0.550	0.341	-1.115	-0.233	-0.842	-1.172	-0.218	1.769	-0.480	-0.345	0.741	0.939	-1.913	1.509	-0.426	0.788	-2.778	U00096	Escherichia coli K-12
Fratul	Gamma Pr	0.562	0.153	0.072	0.196	0.142	0.064	-1.074	0.104	-0.096	-0.383	-0.507	0.570	-0.169	0.703	-0.109	0.883	-0.343	0.447	-0.453	0.836	-1.091	AJ749949	Francisella tularensis
Haeduc	Gamma Pr	0.937	0.288	0.131	0.208	0.310	0.049	-0.713	0.027	0.213	-0.400	0.219	1.342	-0.248	1.497	0.278	1.177	-0.349	0.848	0.225	0.998	-1.562	AE017143	Haemophilus ducreyi
Haeinf	Gamma Pr	1.492	0.328	0.187	0.377	0.600	-0.001	-1.363	0.610	0.650	-0.898	-0.124	1.359	-0.361	0.657	0.577	1.270	-0.939	0.870	0.139	1.139	-2.078	L42023	Haemophilus influenzae
Idiloi	Gamma Pr	1.152	0.248	0.101	0.254	0.550	0.430	-0.604	0.616	0.313	-0.441	0.084	0.725	-0.187	1.162	0.271	1.106	-0.026	0.949	0.355	0.652	-1.299	AE017340	Idiomarina loihensis L2TR
Legpne	Gamma Pr	0.101	0.090	0.036	0.131	-0.156	0.122	-0.302	0.465	0.131	-0.139	-0.409	0.766	0.070	0.742	0.591	0.796	0.158	0.646	0.084	0.685	-0.519	AE017354	Legionella pneumophila Philadelphia 1
Mansuc	Gamma Pr	1.192	0.275	0.138	0.268	0.512	0.302	-1.078	0.900	1.654	-0.489	0.170	1.410	-0.485	1.157	0.330	1.694	-0.027	1.224	0.363	1.119	-2.061	AE016827	Mannheimia succiniciproducens
Metcap	Gamma Pr	-0.265	-0.030	-0.047	-0.028	-0.160	-0.101	-0.121	0.685	-0.351	0.202	-0.224	0.356	-0.090	0.126	0.162	0.171	0.080	0.311	0.440	0.300	-0.062	AE017282	Methylococcus capsulatus
Pasmul	Gamma Pr	1.339	0.311	0.168	0.339	0.521	-0.024	-1.446	-0.096	0.818	-0.612	-0.007	1.413	-0.340	1.009	0.919	1.266	-0.397	0.958	0.298	1.306	-2.039	AE004439	Pasturella multocida
Pholum	Gamma Pr	1.034	0.261	0.184	0.275	0.314	0.209	-0.985	0.480	-0.412	-0.900	-0.246	0.970	-0.316	0.929	0.399	0.465	-1.436	0.738	0.049	0.449	-2.541	BX470251	Photohabdus luminescens

		Strength of Selected Codon Usage Bias for Each Amino Acid																							
Bacterium	Group	WWY	Phe	Tyr	Asn	Ile	Glu	His	Cys	Gln	Asp	Lys	Val AU/GC	Val AG/UC	Pro AU/GC	Pro AG/UC	Thr AU/GC	Thr AG/UC	Ala AU/GC	Ala AG/UC	Gly AU/GC	Gly AG/UC	Accession Number	Species
Phopro	Gamma Pr	1.535	0.345	0.271	0.395	0.525	0.128	-1.629	0.694	-0.534	-0.966	-0.425	1.520	-0.418	0.660	0.385	1.607	-0.976	1.163	-0.289	0.475	-1.950	CR354531*	Photobacterium profundum
Pseaer	Gamma Pr	-0.019	-0.026	0.073	0.141	-0.207	0.616	-0.171	-1.358	0.311	0.501	0.802	1.399	-0.423	0.442	0.216	0.945	-1.346	1.813	-0.622	1.214	-1.632	AE004091	Pseudomonas aeruginosa
Pseflu	Gamma Pr	0.452	0.213	0.092	0.181	-0.034	0.589	-0.285	-0.754	0.942	0.130	0.942	2.108	-0.932	1.874	0.705	1.706	-1.148	2.332	-0.239	1.414	-2.510	CP000076	Pseudomonas fluorescens Pf-5
Pseput	Gamma Pr	0.917	0.316	0.241	0.243	0.117	0.473	-1.017	-0.833	0.190	-0.375	0.507	1.516	-1.147	1.240	0.373	0.660	-1.726	1.385	-0.629	1.240	-2.349	AE015451	Pseudomonas putida
Psesyr	Gamma Pr	0.701	0.202	0.099	0.203	0.197	0.303	-0.621	-0.612	0.345	-0.179	0.200	1.289	-0.628	1.027	0.437	0.977	-1.600	1.433	-0.635	0.842	-2.311	AE016853	Pseudomonas syringae
Psyarc	Gamma Pr	1.037	0.203	0.149	0.338	0.347	0.888	-0.751	0.013	0.273	-0.688	0.211	1.469	-0.278	1.194	0.309	0.663	-0.766	0.927	-0.232	0.454	-1.331	CP000082	Psychrobacter arcticum
Salent	Gamma Pr	1.522	0.334	0.247	0.419	0.522	0.440	-1.310	-0.327	-0.470	-0.925	-0.082	1.690	-0.595	-0.081	0.952	1.169	-1.981	1.797	-0.422	1.011	-2.724	AE006468	Salmonella enterica
Sheone	Gamma Pr	1.377	0.413	0.219	0.385	0.360	0.508	-1.202	0.266	0.205	-1.223	-0.167	1.774	-0.666	2.030	0.855	1.294	-0.842	1.614	-0.255	0.790	-2.270	AE014299	Shewenella oneidensis
Vibcho	Gamma Pr	1.725	0.435	0.355	0.439	0.496	0.217	-1.534	0.684	0.632	-1.333	-0.467	2.371	-0.598	2.371	0.714	1.999	-0.838	1.836	-0.186	0.699	-3.104	AE003852*	Vibrio cholerae
Vibfis	Gamma Pr	2.001	0.395	0.344	0.564	0.698	-0.044	-1.846	0.414	0.357	-1.294	-0.518	1.438	-0.284	0.604	0.328	1.611	-0.988	1.008	-0.526	0.104	-3.033	CP000020*	Vibrio fischeri
Vibpar	Gamma Pr	1.886	0.367	0.482	0.440	0.598	0.192	-1.427	0.807	0.082	-0.885	-0.184	2.338	-0.441	1.903	-0.016	2.459	-1.094	1.885	-0.733	0.604	-3.273	BA000031*	Vibrio parahaemolyticus
Vibvul	Gamma Pr	1.950	0.394	0.389	0.506	0.661	0.265	-1.447	1.427	0.256	-0.955	-0.303	2.873	-0.626	2.041	0.311	2.369	-1.022	1.989	-0.391	0.967	-3.301	AE016795*	Vibrio vulnificus CMCP6
Wigbre	Gamma Pr	0.105	-0.036	-0.005	0.103	0.044	0.026	-0.274	0.564	0.917	0.098	0.003	0.078	0.105	0.159	0.082	-0.064	0.109	-0.242	0.009	-0.180	-0.047	BA000021	Wigglesworthia glossinidia
Xanaxo	Gamma Pr	0.636	0.258	0.073	0.159	0.146	0.230	-0.853	-0.102	-1.093	-0.588	-1.372	0.368	-0.980	-0.723	0.714	-0.319	-0.662	0.446	-0.617	0.785	-2.097	AE008923	Xanthomonas axonopodis
Xancam	Gamma Pr	0.607	0.174	0.070	0.110	0.252	0.227	-0.862	0.253	-0.967	-0.450	-1.334	0.380	-1.171	-0.558	0.595	-0.224	-0.615	0.655	-0.482	0.937	-1.994	AE008922	Xanthomonas campestris
Xanory	Gamma Pr	0.535	0.209	0.071	0.126	0.129	0.332	-0.781	0.142	-0.998	-0.514	-1.408	0.025	-0.914	-0.856	0.690	-0.431	-0.723	0.332	-0.566	0.594	-2.111	AE013598	Xanthomonas oryzae
Xylfas	Gamma Pr	-0.781	-0.091	-0.117	-0.199	-0.375	-0.017	0.864	1.329	-0.308	0.730	-0.316	0.691	-0.616	0.737	-0.655	1.095	-0.557	0.623	-0.659	0.871	-0.882	AE009442	Xylella fastidiosa Temecula
Yerpes	Gamma Pr	1.153	0.270	0.169	0.320	0.393	0.542	-0.615	0.081	-0.063	-0.968	-0.422	1.332	-0.509	0.368	0.660	0.708	-1.338	1.194	-0.152	0.691	-2.339	AL590842	Yersinia pestis CO92
Bdebac	Delta Pr	1.060	0.379	0.215	0.148	0.318	-0.058	-0.749	-0.138	1.228	-0.221	0.438	2.458	-0.733	2.075	-0.028	2.434	-0.308	1.971	-0.004	0.976	-1.159	BX842601	Bdellovibrio bacteriovorus
Despsy	Delta Pr	0.056	0.078	0.022	0.024	-0.068	-0.040	-0.264	0.843	0.594	-0.160	-0.266	1.481	-0.208	1.432	0.299	0.561	-0.569	1.390	0.186	1.147	-0.398	CR522870	Desulfotalea psychrophila
Desvul	Delta Pr	0.473	0.176	0.127	0.153	0.018	0.485	-0.415	-1.647	-1.598	-0.050	-0.559	0.471	-0.543	0.098	-0.587	-0.454	-1.187	0.232	-1.023	0.521	-1.563	AE017285	Desulfovibrio vulgaris
Geosul	Delta Pr	-0.384	0.036	-0.106	-0.095	-0.219	0.175	0.125	0.253	0.192	0.169	-0.228	0.670	-0.495	0.527	-0.097	0.462	-0.034	0.618	0.100	0.567	-0.487	AE017180	Geobacter sulfurreducens
Camjej	Epsilon Pr	0.486	0.189	0.043	0.097	0.157	0.242	-0.937	-0.588	0.338	-0.343	0.272	0.763	-0.188	0.142	1.143	1.154	0.050	0.628	0.288	0.460	-0.861	AL111168	Campylobacter jejuni 11168
Helhep	Epsilon Pr	0.019	0.006	0.089	-0.038	-0.038	0.034	-0.236	-0.275	-0.257	-0.128	-0.255	0.435	-0.108	0.208	0.407	0.109	0.231	0.240	0.139	0.387	-0.238	AE017125	Helicobacter hepaticus
Helpyl	Epsilon Pr	0.016	0.026	0.021	-0.039	0.009	0.084	-0.099	0.099	-0.354	-0.212	-0.162	-0.072	0.304	0.356	0.494	0.069	-0.247	0.295	0.112	0.400	-0.494	AE001439	Helicobacter pylori J99
Wolsuc	Epsilon Pr	0.563	0.183	0.024	0.208	0.148	0.723	-0.480	-1.232	0.183	-0.352	0.167	1.205	-0.525	0.582	-0.632	0.682	-0.338	0.696	-0.002	0.512	-1.185	BX571656	Wolinella succinogenes
Bacant	Firmicutes	2.045	0.435	0.322	0.537	0.751	0.027	-1.541	-0.767	0.888	-1.362	0.321	1.181	-0.215	2.049	-0.413	1.743	-1.222	1.471	-0.867	0.585	-1.361	AE016879	Bacillus anthracis Ames
Baccla	Firmicutes	0.767	0.172	0.145	0.289	0.161	0.187	-1.289	-0.151	1.357	-0.461	0.392	0.788	0.037	0.576	0.443	0.786	-0.079	0.442	0.216	0.442	-0.830	AP006627	Bacillus clausii
Bachal	Firmicutes	0.999	0.251	0.208	0.352	0.188	0.337	-1.099	0.111	1.109	-0.487	-0.014	0.774	0.095	1.273	0.064	1.123	-0.055	0.595	0.271	0.931	-0.838	BA000004	Bacillus halodurans
Baclic	Firmicutes	1.072	0.265	0.222	0.358	0.226	0.278	-0.864	0.127	1.323	-0.370	0.648	1.540	0.094	1.890	-0.246	1.780	-0.717	1.357	-0.263	1.397	-0.554	CP000002	Bacillus licheniformis strain ATCC 14580
Bacsub	Firmicutes	1.360	0.326	0.296	0.402	0.336	0.470	-1.273	0.539	1.337	-0.453	0.829	1.751	-0.225	2.016	-0.261	1.627	-0.984	1.451	-0.500	1.431	-0.669	AL009126	Bacillus subtilis
Cloace	Firmicutes	0.838	0.265	0.130	0.234	0.209	0.728	-0.588	-0.570	-0.089	-0.144	0.228	1.158	-0.379	1.798	0.529	2.206	0.192	1.218	0.112	0.774	-0.194	AE001437	Clostridium acetobutylicum
Cloper	Firmicutes	2.648	0.584	0.474	0.774	0.815	0.249	-1.782	-0.168	1.666	-1.423	0.190	1.117	-0.422	1.903	0.870	3.125	-0.178	1.370	-0.262	0.955	-0.514	BA000016	Clostridium perfringens
Clotet	Firmicutes	1.004	0.264	0.159	0.315	0.266	0.706	-1.248	-0.018	1.234	-0.630	-0.326	1.035	-0.389	1.284	1.223	3.123	0.222	0.644	0.096	0.428	-0.216	AE015927	Clostridium tetani
Entfae	Firmicutes	1.840	0.283	0.256	0.580	0.721	1.279	-1.222	-1.110	2.078	-0.922	1.098	1.269	0.044	1.938	0.114	2.036	-0.378	1.016	-0.079	0.766	-0.449	AE016830	Enterococcus faecalis
Geokau	Firmicutes	0.559	0.192	0.097	0.143	0.127	0.706	-0.170	-1.034	1.499	-0.478	1.018	0.489	-0.147	0.014	1.062	-0.216	0.524	0.535	0.389	0.183	-0.744	BA000043	Geobacillus kaustophilus
Lacaci	Firmicutes	1.361	0.153	0.371	0.485	0.352	1.595	-1.869	-1.258	3.672	-1.221	-1.675	1.343	-0.628	2.043	0.129	0.856	-1.501	0.961	-0.551	0.505	-2.071	CP000033	Lactobacillus acidophilus
Lacjoh	Firmicutes	1.502	0.198	0.407	0.510	0.387	1.745	-1.925	-1.738	3.112	-0.846	-0.972	0.990	-0.723	2.660	-0.056	0.882	-1.559	1.268	-0.239	0.701	-2.328	AE017198	Lactobacillus johnsonii
Laclac	Firmicutes	2.288	0.368	0.369	0.717	0.834	2.208	-2.114	-2.674	4.961	-0.877	1.212	0.992	-0.037	2.966	0.377	2.703	-0.314	0.967	-0.004	1.085	-1.124	AE005176	Lactococcus lactis lactis
Lacpla	Firmicutes	1.253	0.213	0.195	0.426	0.418	2.351	-0.210	-0.310	2.454	-0.495	-0.548	1.507	-0.869	2.536	-0.719	1.053	-0.913	0.878	-0.798	0.816	-1.253	AL935263	Lactobacillus plantarum
Lismon	Firmicutes	1.198	0.240	0.156	0.339	0.464	0.585	-0.886	0.614	2.851	-0.707	0.538	1.028	-0.003	1.380	-0.479	1.858	-1.022	1.125	-0.585	0.736	-0.687	AL591824	Listeria monocytogenes EGD
Mesflo	Firmicutes	1.418	0.227	0.203	0.497	0.491	0.699	-1.493	0.294	3.960	-0.831	0.761	0.982	-0.109	1.054	0.188	1.688	-0.203	0.771	-0.137	-0.184	0.537	AE017263	Mesoplasma florum
Mycgal	Firmicutes	0.498	0.104	0.138	0.275	-0.019	0.698	-0.893	-0.295	0.607	-0.760	0.010	0.492	0.118	0.886	-0.098	0.605	-0.131	0.304	-0.126	0.508	-0.485	AE015450	Mycoplasma gallisepticum
Mycgen	Firmicutes	0.318	0.001	0.034	0.139	0.145	-0.290	-0.571	-0.400	0.095	0.113	-0.176	-0.237	0.120	-0.338	-0.132	-0.342	-0.305	-0.159	0.048	-0.205	-0.033	L43967	Mycoplasma genitalium
Mychyo	Firmicutes	0.101	0.060	0.000	0.023	0.019	0.078	-0.342	-0.038	-0.051	0.086	-0.091	0.059	-0.082	-0.034	0.011	0.027	0.004	-0.146	0.133	0.234	0.276	AE017332	Mycoplasma hyopneumoniae strain 232
Mycmob	Firmicutes	0.434	0.092	0.028	0.156	0.158	0.279	-0.771	0.369	0.236	-0.162	0.225	0.718	-0.160	0.708	0.283	0.602	0.207	0.090	-0.034	0.332	-0.252	AE017308	Mycoplasma mobile
Mycmyc	Firmicutes	0.650	0.165	0.140	0.175	0.171	0.916	-0.899	-1.533	0.993	-0.261	0.647	1.073	-0.160	3.073	0.252	0.687	-0.153	0.781	-0.013	0.664	0.187	BX293980	Mycoplasma mycoides
Mycpen	Firmicutes	0.496	0.086	0.106	0.234	0.069	1.112	-0.712	0.528	1.335	-0.433	0.142	0.592	-0.034	0.534	-0.136	0.857	-0.190	0.793	-0.314	0.323	-0.277	BA000026	Mycoplasma penetrans
Mycpne	Firmicutes	0.324	0.065	0.056	0.211	-0.008	0.214	-0.749	-0.204	0.448	-0.287	-0.249	0.191	0.341	0.105	0.245	-0.296	-0.276	0.079	0.046	0.285	-0.129	U00089	Mycoplasma pneumoniae

Appendix C

		Strength of Selected Codon Usage Bias for Each Amino Acid																							
Bacterium	Group	WWY	Phe	Tyr	Asn	Ile	Glu	His	Cys	Gln	Asp	Lys	Val AU/GC	Val AG/UC	Pro AU/GC	Pro AG/UC	Thr AU/GC	Thr AG/UC	Ala AU/GC	Ala AG/UC	Gly AU/GC	Gly AG/UC	Accession Number	Species
Mycpul	Firmicutes	0.380	0.123	0.087	0.150	0.020	0.329	-0.639	-0.024	0.419	-0.292	0.225	0.481	-0.205	1.106	0.183	0.194	0.123	0.496	0.021	0.779	-0.335	AL445566	Mycoplasma pulmonis
Mycsyn	Firmicutes	0.636	0.153	0.124	0.201	0.158	0.469	-0.734	0.866	0.916	-0.697	0.188	0.568	-0.331	1.760	-0.074	-0.231	0.209	0.294	-0.046	0.779	-0.109	AE017245	Mycoplasma synoviae
Oceihe	Firmicutes	1.301	0.255	0.205	0.447	0.393	0.133	-1.302	-1.120	0.644	-0.722	0.536	0.817	-0.139	1.007	0.457	1.302	-0.547	0.558	-0.180	0.458	-0.820	BA000028	Oceanobacillus iheyensis
Phyast	Firmicutes	0.218	0.071	0.099	0.106	-0.058	-0.033	-0.471	-0.468	-0.551	-0.540	0.379	0.391	0.171	0.467	0.320	0.348	-0.237	0.252	-0.095	-0.052	-0.108	AP006628	Phytoplasma asteris OY
Staaur	Firmicutes	1.564	0.318	0.225	0.457	0.564	0.623	-1.793	-0.002	2.260	-1.248	1.141	1.777	0.146	1.759	0.188	1.672	-0.890	1.262	-0.682	0.972	-0.533	BA000018	Staphylococcus aureus N315
Staepi	Firmicutes	1.164	0.284	0.161	0.345	0.374	1.065	-1.427	-0.165	2.339	-1.026	1.096	1.126	-0.023	1.817	0.483	1.842	-0.563	1.074	-0.295	0.699	-0.792	AE015929	Staphylococcus epididermis
Stahae	Firmicutes	1.442	0.301	0.179	0.435	0.527	1.048	-1.594	-0.544	3.294	-0.947	1.566	1.563	-0.042	1.862	0.398	2.395	-0.830	1.314	-0.571	0.620	-0.708	AP006716	Staphylococcus haemolyticus
Stasap	Firmicutes	1.355	0.310	0.141	0.478	0.426	0.850	-1.491	0.122	2.370	-0.967	1.039	1.612	-0.174	1.679	0.205	2.094	-0.818	1.223	-0.565	0.798	-0.429	AP008934	Staphylococcus saprophyticus
Straga	Firmicutes	1.504	0.245	0.219	0.510	0.529	0.969	-1.746	-2.091	1.479	-0.812	1.367	1.142	-0.305	2.640	0.642	2.212	-0.318	0.899	0.096	1.148	-0.700	AE009948	Streptococcus agalactiae 2603V/R
Strpne	Firmicutes	1.720	0.279	0.297	0.559	0.585	1.065	-2.066	0.165	2.982	-0.821	1.844	1.525	-0.092	1.829	0.998	2.473	-0.150	1.382	0.286	1.512	-0.669	AE007317	Streptococcus pneumoniae R6
Strpyo	Firmicutes	1.759	0.330	0.275	0.497	0.657	1.083	-1.487	-1.553	1.866	-0.748	1.458	1.396	-0.363	2.289	0.776	1.913	-0.044	1.358	0.268	0.977	-0.753	AE004092	Streptococcus pyogenes M1 GAS SF370
Strthe	Firmicutes	1.656	0.245	0.284	0.568	0.559	1.491	-1.969	-1.090	2.493	-0.546	1.713	1.480	-0.192	3.537	0.761	2.298	-0.349	1.074	0.263	0.909	-0.687	CP000023	Streptococcus thermophilus LMG 18311
Symthe	Firmicutes	-0.150	0.059	-0.006	-0.053	-0.150	-0.330	0.149	-1.835	-0.707	0.187	-1.163	0.217	-0.591	-0.036	-0.332	-0.448	-0.426	0.167	-0.319	0.168	-1.270	AP006840	Symbiobacterium thermophilum
Theten	Firmicutes	0.457	0.162	0.016	0.152	0.126	-0.368	-0.347	-1.047	-0.669	-0.289	-0.357	0.157	-0.015	0.302	0.202	-0.249	0.137	0.032	0.143	0.151	-0.544	AE008691	Thermoanaerobacter tengcongensis
Ureure	Firmicutes	0.401	0.076	0.148	0.198	-0.020	0.782	-0.596	-0.267	1.952	-0.223	-0.052	0.384	0.220	1.412	0.411	0.838	0.085	0.459	0.011	0.211	-0.159	AF222894	Ureaplasma urealyticum
Biflon	Actinobacteria	1.343	0.341	0.363	0.352	0.286	-0.491	-1.766	-1.570	-2.277	-0.448	-3.108	0.230	-0.996	-0.531	0.375	-0.199	-1.492	0.793	-1.236	0.044	-1.848	AE014295	Bifidobacterium longum
Cordip	Actinobacteria	1.861	0.394	0.442	0.568	0.458	-1.107	-2.586	-0.927	-2.630	-0.920	-2.872	0.686	-1.375	1.571	0.645	-1.499	-2.238	1.493	-0.631	-0.097	-2.165	BX248353	Corynebacterium dipheriae
Coreff	Actinobacteria	1.039	0.114	0.472	0.424	0.028	-0.441	-1.481	-0.980	-1.295	-0.794	-2.909	1.069	-1.505	1.236	0.589	-0.661	-1.649	1.679	-0.320	-0.048	-1.732	BA000035	Corynebacterium efficiens
Corglu	Actinobacteria	2.185	0.435	0.669	0.609	0.471	-1.045	-1.990	-0.549	-3.160	-1.179	-3.290	0.798	-1.204	1.653	0.241	-1.129	-2.410	1.452	-0.311	-0.205	-1.413	BA000036	Corynebacterium glutamicum
Corjej	Actinobacteria	1.588	0.437	0.476	0.478	0.197	-0.820	-2.495	0.467	-2.917	-0.701	-4.070	0.862	-1.269	-0.384	0.813	-0.731	-1.603	1.488	-0.506	0.234	-2.131	CR931997	Corynebacterium jejkeium
Leixyl	Actinobacteria	0.522	0.116	0.194	0.237	-0.025	-1.025	-1.287	1.029	-2.004	-1.024	-1.238	0.127	-0.706	-1.033	0.281	-0.568	-0.826	0.281	-0.756	0.359	-1.303	AE016822	Leifsonia xyli
Mycavi	Actinobacteria	1.184	0.396	0.208	0.410	0.169	-0.867	-1.615	-0.818	-0.566	-0.986	-1.466	-0.696	-0.391	-0.341	-0.058	-0.386	-0.363	-0.136	-0.257	0.339	-0.896	AE016958	Mycobacterium avium
Myclep	Actinobacteria	0.515	0.195	0.088	0.151	0.081	-0.600	-0.421	-2.118	-0.978	-0.358	-0.759	-0.342	-0.267	-0.367	-0.028	-0.183	-0.483	-0.327	-0.225	-0.011	-0.921	AL450380	Mycobacterium leprae
Myctub	Actinobacteria	0.453	0.115	0.104	0.132	0.102	-0.533	-0.251	-0.510	-0.798	-0.590	-0.844	-0.133	-0.149	-0.025	-0.018	-0.281	-0.311	-0.063	-0.156	-0.060	-0.636	AL123456	Mycobacterium tuberculosis
Nocfar	Actinobacteria	1.413	0.495	0.405	0.336	0.177	-0.885	-1.321	-0.404	-2.462	-0.418	-2.608	-0.048	-0.643	0.032	0.070	-0.334	-0.622	0.146	-0.364	0.518	-1.804	AP006618	Nocardia farcinica
Proacn	Actinobacteria	0.621	0.196	0.203	0.174	0.048	-1.526	-0.744	-0.659	-2.335	-0.515	-2.000	-0.279	-0.421	-0.578	-0.026	-0.266	-1.217	0.091	-1.004	0.131	-1.209	AE017283	Propionibacterium acnes
Strave	Actinobacteria	0.686	0.061	0.337	0.465	-0.177	-1.558	-3.114	0.423	-2.741	-1.078	-3.599	0.437	-0.835	-0.197	0.213	-0.008	-0.383	0.593	-0.621	0.931	-2.101	BA000030	Streptomyces avermitilis
Strcoe	Actinobacteria	0.987	0.180	0.559	0.129	0.120	-1.148	-2.405	0.222	-3.888	-1.235	-4.778	0.310	-0.647	-0.294	0.120	-0.117	-0.230	0.333	-0.467	0.962	-1.674	AL645882	Streptomyces coelicolor
Thefus	Actinobacteria	0.439	-0.020	0.291	0.396	-0.229	-1.184	-2.186	-0.796	-2.443	-0.739	-2.135	0.254	-0.452	-0.184	0.369	0.004	-0.019	0.154	-0.296	0.363	-1.181	CP000088	Thermobifida fusca
Trowhi	Actinobacteria	0.014	-0.003	0.037	0.000	-0.019	-0.533	0.223	-0.135	-0.936	0.078	-0.588	-0.022	-0.306	-0.258	-0.299	-0.181	-0.284	-0.248	-0.261	0.093	-0.475	AE014184	Tropheryma whipplei Twist
Glovio	Cyanobacteria	0.370	0.110	0.049	0.109	0.102	-0.101	-0.534	-0.805	-0.345	-0.223	-0.437	-0.680	-0.438	-0.406	-0.509	-0.832	-0.277	-0.479	-0.352	-0.286	-0.949	BA000045	Gloeobacter violaceous
Nostoc	Cyanobacteria	0.763	0.122	0.177	0.308	0.156	0.270	-0.799	-1.301	0.224	-0.701	-0.060	0.305	0.020	0.300	-0.089	-0.317	-0.243	0.157	-0.011	0.169	-0.746	BA000019	Nostoc sp. PCC7120
Pro137	Cyanobacteria	0.044	0.037	0.011	0.003	-0.007	-0.287	-0.388	0.163	-0.513	0.021	-0.471	-0.165	-0.175	0.050	-0.152	-0.150	-0.345	-0.193	-0.363	-0.096	-0.290	AE017126	Prochlorococcus marinus marinus CCMP1375
Promed	Cyanobacteria	0.445	0.100	0.058	0.147	0.140	-0.222	-0.353	0.132	-0.001	-0.234	-0.103	0.058	-0.167	-0.009	-0.012	-0.101	-0.068	-0.130	-0.237	0.196	-0.391	BX548174	Prochlorococcus marinus pastoris CCMP1986 MED4
Promit	Cyanobacteria	0.715	0.205	0.134	0.157	0.220	0.189	-0.896	-0.520	-0.137	-0.912	-0.353	-0.247	-0.251	-0.368	-0.392	-0.541	-0.474	-0.390	-0.439	-0.243	-0.585	BX549175	Prochlorococcus marinus strain MIT9313
Pronat	Cyanobacteria	0.433	0.113	0.082	0.154	0.084	-0.159	-0.778	-0.444	-0.203	-0.265	-0.299	0.100	-0.283	0.142	-0.272	-0.298	-0.065	-0.154	-0.281	0.034	-0.318	CP000095	Prochlorococcus marinus strain NATL2A
Sy6803	Cyanobacteria	0.616	0.085	0.098	0.180	0.253	0.350	-0.943	0.417	0.191	-0.375	-0.231	0.194	-0.179	0.059	-0.734	-0.504	-0.769	0.126	-0.439	0.299	-0.677	BA000022	Synechocystis PCC6803
Synelo	Cyanobacteria	0.776	0.123	0.175	0.253	0.224	0.330	-1.309	-0.746	0.451	-0.896	0.055	0.095	-0.288	-0.506	-0.193	-0.651	-0.449	0.341	-0.403	0.324	-1.084	AP008231	Synechococcus elongatus PCC6301
Synspp	Cyanobacteria	0.918	0.314	0.154	0.216	0.234	0.198	-1.345	-0.829	-0.681	-0.778	-0.695	-0.182	-0.320	-0.444	-0.789	-1.158	-0.891	-0.192	-0.525	-0.126	-1.194	BX548020	Synechococcus sp. WH8102
Theelo	Cyanobacteria	0.178	0.062	0.070	0.091	-0.045	0.254	-0.552	-1.129	0.510	-0.250	0.423	0.165	-0.182	0.066	-0.334	-0.060	-0.356	0.037	-0.383	0.376	-0.619	BA000039	Thermosynechococcus elongatus
Aquaeo	Others	0.393	0.130	0.009	0.164	0.091	0.226	-1.098	-0.530	0.187	-0.392	-0.479	0.092	-0.260	-0.413	-0.265	-0.093	-0.043	0.268	0.105	0.616	-0.263	AE000657	Aquifex aeolicus
Bacfra	Others	0.383	0.147	0.055	0.176	0.005	0.839	-1.269	0.017	-0.642	-0.177	-0.150	1.556	-0.410	0.334	0.243	1.625	-0.551	1.735	-0.353	1.127	-1.244	AP006841	Bacteroides fragilis
Bacthe	Others	0.237	0.051	0.057	0.191	-0.062	0.745	-1.195	0.006	0.209	-0.027	-0.127	1.495	-0.299	0.627	0.106	1.298	-0.460	1.593	-0.513	1.004	-1.171	AE015928	Bacteroides thetaiotamicron
Borbur	Others	-0.308	-0.058	-0.127	-0.099	-0.023	-0.099	0.075	0.774	-0.609	0.023	-0.466	0.336	-0.214	0.367	0.066	0.004	-0.481	0.183	-0.268	0.147	-0.311	AE000783	Borrelia burgdorferi
Borgar	Others	-0.206	-0.007	-0.143	-0.115	0.058	-0.195	0.229	0.440	-0.599	-0.015	-0.512	0.370	-0.224	0.325	0.194	0.100	-0.382	0.177	-0.347	0.021	-0.336	CP000013	Borrelia garinii
Chlabo	Others	-0.148	-0.032	0.025	-0.075	-0.066	0.231	-0.483	-0.171	0.041	-0.265	-0.124	0.288	-0.244	0.807	-0.051	0.295	-0.132	0.609	-0.269	0.135	-0.436	CR848038	Chlamydophila abortus
Chlcav	Others	0.113	0.028	0.046	0.033	0.006	0.311	-0.653	-0.203	-0.004	-0.069	0.055	0.244	-0.278	1.069	0.119	0.268	-0.298	0.280	-0.270	0.143	-0.552	AE015925	Chlamydophila caviae
Chlmur	Others	0.145	0.076	0.055	0.058	-0.044	-0.194	-0.389	-0.079	-0.328	-0.021	-0.308	0.323	-0.314	1.228	-0.052	0.318	-0.058	0.157	-0.298	0.166	-0.425	AE002160	Chlamydia muridarum
Chlpne	Others	-0.065	0.006	0.047	-0.015	-0.103	0.778	-0.044	-0.262	0.216	-0.036	0.277	0.465	-0.130	0.950	-0.172	0.537	-0.187	0.443	-0.164	0.334	-0.356	AE002161	Chlamydophila pneumoniae AR39

		Strength of Selected Codon Usage Bias for Each Amino Acid																						
Bacterium	**Group**	**WWY**	**Phe**	**Tyr**	**Asn**	**Ile**	**Glu**	**His**	**Cys**	**Gln**	**Asp**	**Lys**	**Val AU/GC**	**Val AG/UC**	**Pro AU/GC**	**Pro AG/UC**	**Thr AU/GC**	**Thr AG/UC**	**Ala AU/GC**	**Ala AG/UC**	**Gly AU/GC**	**Gly AG/UC**	**Accession Number**	**Species**
Chltep	Others	0.069	0.080	-0.037	0.020	0.006	-0.031	0.182	-0.639	-0.782	0.606	-0.248	0.457	-0.539	0.387	-0.149	-0.100	-0.500	0.885	-0.465	0.696	-0.330	AE006470	Chlorobium tepidum
Chltra	Others	0.132	0.052	0.056	0.043	-0.020	-0.237	-0.411	-0.288	-0.089	-0.122	-0.291	0.407	-0.416	1.026	-0.149	0.251	-0.156	0.342	-0.250	0.153	-0.369	AE001273	Chlamydia trachomatis
Deheth	Others	0.063	0.094	-0.004	-0.030	0.003	0.023	0.339	0.186	-1.223	-0.106	-0.559	0.602	-0.229	0.325	-0.673	0.162	-0.700	0.313	-0.529	0.436	-1.175	CP000027	Dehalococcoides ethenogenes
Deirad	Others	1.491	0.297	0.230	0.492	0.471	0.750	-1.959	-2.209	-0.630	-0.452	-1.284	-0.169	-0.345	-0.478	-1.185	-0.819	-1.230	0.065	-0.845	0.522	-1.341	AE000513*	Deinococcus radiodurans
Fusnuc	Others	1.244	0.261	0.201	0.419	0.362	1.345	-0.800	0.407	1.705	-0.680	0.032	0.587	0.129	1.774	0.015	1.487	-0.224	0.760	-0.177	0.323	0.250	AE009951	Fusobacterium nucleatum
Lepint	Others	0.670	0.217	0.064	0.154	0.235	-0.268	-0.791	-0.855	-0.531	-0.428	-0.318	-0.060	0.187	0.043	-0.083	0.106	-0.596	0.146	0.019	-0.150	-0.283	AE010300*	Leptospira interrogans Lai
Parspp	Others	0.347	0.094	0.044	0.152	0.056	0.113	-0.830	-0.454	-0.206	-0.341	-0.385	0.185	-0.017	0.791	0.716	0.376	0.309	0.210	0.118	0.341	-0.767	BX908798	Parachlamydia sp. UWE25
Porgin	Others	0.021	0.035	-0.021	0.061	-0.054	0.352	-0.646	0.323	-0.216	-0.106	-0.455	0.543	0.048	0.366	-0.653	0.279	0.092	0.616	-0.023	0.662	-0.917	AE015924	Porphyromonas ginigivalis
Rhobal	Others	0.825	0.197	0.152	0.322	0.154	0.734	-1.501	-0.871	0.284	-0.494	-0.438	0.278	-0.508	1.040	-0.248	-0.194	-0.542	0.517	-0.641	-0.003	-0.978	BX119912	Rhodopirellula baltica
Themar	Others	0.365	0.105	0.120	0.017	0.123	0.282	-0.298	-0.408	-0.659	-0.016	-0.247	0.266	-0.292	0.141	-0.314	0.412	0.072	0.292	-0.300	0.640	-0.090	AE000512	Thermotoga maritima
Thethe	Others	-0.158	-0.061	0.133	-0.039	-0.191	-0.547	-0.733	-1.425	-1.326	-0.516	-1.197	0.147	0.234	-0.148	0.139	-2.591	0.086	0.336	0.220	0.001	-0.759	AE017221	Thermus thermophilus
Treden	Others	0.620	0.190	0.101	0.172	0.157	0.012	-0.743	-0.584	-0.626	-0.086	-0.666	0.445	-0.221	-0.623	-0.502	0.134	-0.115	0.438	-0.064	0.374	-0.370	AE017226	Treponema denticola
Trepal	Others	-0.015	-0.004	-0.014	-0.020	0.023	-0.651	0.429	0.834	-0.272	0.095	-0.739	0.047	-0.105	0.240	-0.051	-0.048	-0.168	-0.092	-0.222	-0.030	-0.111	AE000520	Treponema pallidum

Appendix C

Amino Acid	Met	Trp	Phe	Phe	Tyr	Tyr	Ile	Ile	Ile	Asn	Asn	His	His	Asp	Asp	Cys	Cys	Sel	Gln	Gln	Lys	Lys	Glu	Glu	Pro	Pro
AntiCodon	CAU	CCA	AAA	GAA	AUA	GUA	AAU	GAU	UAU	AUU	GUU	AUG	GUG	AUC	GUC	ACA	GCA	UCA	UUG	CUG	UUU	CUU	UUC	CUC	AGG	GGG
AciAdp	6	2	0	2	0	1	0	7	0	0	4	0	1	0	3	0	1	0	5	0	2	0	5	0	0	0
AgrTum	6	1	0	1	0	1	0	4	0	0	1	0	1	0	2	0	1	0	1	1	1	1	2	0	0	1
AnaMar	3	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1	1	1	0	0	1
AquAeo	3	1	0	1	0	1	0	2	0	0	1	0	1	0	1	0	1	1	1	0	1	1	1	0	0	1
AzoEbn	3	1	0	1	0	1	0	4	0	0	2	0	1	0	2	0	2	0	1	0	1	2	2	0	0	1
BacAnt	8	2	0	4	0	2	0	4	0	0	5	0	2	0	6	0	1	0	4	0	5	0	7	0	0	0
BacCla	6	1	0	2	0	2	0	3	0	0	4	0	2	0	3	0	1	0	3	0	3	0	4	0	0	0
BacFra	3	2	0	2	0	2	0	6	0	0	2	0	1	0	3	0	1	0	1	1	2	2	1	1	0	1
BacHal	6	1	0	2	0	3	0	3	0	0	4	0	2	0	3	0	1	0	4	0	3	0	4	0	0	0
BacLic	6	1	0	3	0	2	0	3	0	0	3	0	2	0	4	0	1	0	4	0	3	0	6	0	0	0
BacSub	6	1	0	3	0	2	0	3	0	0	4	0	2	0	4	0	1	0	4	0	4	0	6	0	0	0
BacThe	3	2	0	2	0	2	0	5	0	0	2	0	1	0	3	0	1	0	1	1	2	2	1	1	0	1
BarHen	5	1	0	1	0	1	0	2	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	0	0	1
BarQui	4	1	0	1	0	1	0	2	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	0	0	1
BdeBac	3	1	0	1	0	1	0	2	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	1
BifLon	3	2	0	1	0	1	0	1	0	0	3	0	1	0	2	0	1	0	1	1	1	1	1	1	0	1
BloFlo	3	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1	1	1	0	0	0
BloPen	3	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1	1	1	0	0	1
BorBur	3	1	0	2	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1	1	0	0	0	0
BorGar	3	1	0	2	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1	1	0	0	0	0
BorPer	4	1	0	1	0	1	0	3	0	0	2	0	1	0	1	0	1	0	1	0	1	1	1	0	0	1
BraJap	5	1	0	2	0	1	0	1	0	0	2	0	1	0	1	0	1	0	1	1	1	1	1	1	0	1
BruMel	5	1	0	1	0	1	0	3	0	0	1	0	1	0	2	0	1	0	1	1	1	1	1	1	0	1
BuchAp	3	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	0	0	1	0	1	0	1	0	0	0
BuchBp	3	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	0	0	1	0	1	0	1	0	0	0
BuchSg	3	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	0	0	1	0	1	0	1	0	0	0
BurPse	4	1	0	1	0	1	0	4	0	0	2	0	1	0	3	0	2	0	1	0	1	1	2	0	0	1
CamJej	3	1	0	1	0	1	0	3	0	0	1	0	1	0	2	0	1	0	1	0	2	0	1	0	0	0
CauCre	5	1	0	1	0	1	0	2	0	0	2	0	1	0	2	0	1	0	1	0	1	1	2	0	0	1
ChlAbo	3	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	1
ChlCav	3	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	1
ChlMur	3	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	1
ChlPne	3	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	1
ChlTep	4	1	0	1	0	1	0	2	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1	0	1
ChlTra	3	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	1
ChrVio	5	1	0	2	0	2	0	8	0	0	4	0	3	0	7	0	1	0	2	0	3	2	4	0	0	1
CloAce	4	1	0	3	0	2	0	0	0	0	4	0	2	0	3	0	2	0	2	1	4	2	2	1	0	0
CloPer	8	2	0	4	0	3	0	4	0	0	4	0	2	0	3	0	2	1	2	0	7	2	3	0	0	0
CloTet	5	1	0	2	0	2	0	3	0	0	3	0	1	0	1	0	2	0	1	1	2	1	1	1	0	1
ColPsy	9	2	0	2	0	2	0	2	0	0	4	0	2	0	5	0	1	0	2	0	6	0	6	0	0	1
CorDip	3	1	0	1	0	1	0	2	0	0	1	0	1	0	2	0	1	0	0	1	1	1	1	2	0	1
CorEff	4	1	0	1	0	1	0	1	0	0	1	0	1	0	2	0	1	0	1	2	1	2	1	3	0	1
CorGlu	4	1	0	1	0	1	0	2	0	0	2	0	1	0	2	0	1	0	1	2	1	2	1	3	0	1
CorJej	3	1	0	1	0	1	0	1	0	0	1	0	1	0	2	0	1	0	1	1	1	1	1	2	0	1
CoxBur	3	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1	1	1	1	0	1
DecAro	4	1	0	1	0	1	0	4	0	0	2	0	1	0	3	0	1	1	1	0	1	1	2	0	0	1
DehEth	3	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1	0	1
DeiRad	3	1	0	1	0	2	0	1	0	0	1	0	1	0	3	0	1	0	1	1	1	1	1	0	0	1
DesPsy	5	1	0	2	0	2	0	2	0	0	1	0	2	0	6	0	1	1	2	0	2	0	2	2	0	1
DesVul	4	1	0	1	0	1	0	5	0	0	2	0	1	0	2	0	1	1	1	1	1	2	3	0	0	1
EhrCan	3	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1	1	1	0	0	1
EhrRum	3	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1	1	1	0	0	1
EntFae	9	3	0	2	0	2	0	2	0	0	2	0	1	0	3	0	2	0	2	0	3	1	3	0	0	0
ErwCar	7	1	0	2	0	2	0	3	0	0	3	0	1	0	3	0	1	0	2	1	3	1	4	0	0	1
EscCol	8	1	0	2	0	3	0	3	0	0	4	0	1	0	3	0	1	1	2	2	6	0	4	0	0	1
FraTul	3	1	0	1	0	1	0	3	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	0
FusNuc	3	1	0	2	0	1	0	1	0	0	2	0	1	0	3	0	1	0	2	0	2	1	3	0	0	0
GeoKau	5	1	0	1	0	2	0	4	0	0	4	0	2	0	3	0	1	0	4	0	4	0	5	0	0	1
GeoSul	3	1	0	1	0	1	0	2	0	0	1	0	1	0	2	0	1	1	2	0	1	0	2	0	0	1
GloVio	2	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1	0	1
GluOxy	6	1	0	1	0	1	0	4	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1	0	1

Amino Acid	Pro	Pro	Thr	Thr	Thr	Thr	Val	Val	Val	Val	Ala	Ala	Ala	Ala	Gly	Gly	Gly	Gly	Leu	Leu	Leu	Leu	Leu	Leu	Ser	Ser
AntiCodon	UGG	CGG	AGU	GGU	UGU	CGU	AAC	GAC	UAC	CAC	AGC	GGC	UGC	CGC	ACC	GCC	UCC	CCC	UAA	CAA	AAG	GAG	UAG	CAG	AGA	GGA
AciAdp	2	0	0	1	2	0	0	1	3	0	0	1	7	0	0	3	1	0	2	1	0	1	2	0	0	1
AgrTum	1	1	0	1	1	1	0	1	1	0	0	1	4	0	0	2	1	1	1	1	0	1	1	1	0	1
AnaMar	1	0	0	1	1	0	0	1	1	0	0	0	1	0	0	1	1	1	1	1	0	1	1	1	0	1
AquAeo	1	1	0	1	1	1	0	1	1	0	0	1	2	0	0	1	1	1	1	1	0	1	1	1	0	1
AzoEbn	1	1	0	1	1	1	0	1	1	2	0	2	4	1	0	2	1	1	1	1	0	1	1	2	0	1
BacAnt	3	0	0	1	4	0	0	1	5	0	0	0	5	0	0	4	4	0	2	1	0	1	2	0	0	1
BacCla	2	0	0	1	2	1	0	1	4	0	0	1	4	0	0	3	3	0	2	1	0	1	2	1	0	1
BacFra	2	1	0	1	3	1	0	0	3	0	0	1	6	0	0	4	2	1	1	1	0	1	1	2	0	1
BacHal	2	0	0	1	4	0	0	1	4	0	0	1	5	0	0	4	3	0	2	1	0	1	2	0	0	1
BacLic	1	0	0	1	2	1	0	1	3	0	0	1	3	0	0	2	3	0	1	1	0	1	1	1	0	1
BacSub	3	0	0	1	4	0	0	1	4	0	0	1	5	0	0	4	3	0	3	1	0	1	2	1	0	1
BacThe	2	1	0	1	3	1	0	0	2	0	0	1	5	0	0	4	2	1	1	1	0	1	1	2	0	2
BarHen	1	0	0	1	1	1	0	1	1	0	0	1	2	0	0	1	1	0	1	1	0	1	1	1	0	1
BarQui	1	0	0	1	1	1	0	1	1	0	0	1	2	0	0	1	1	0	1	1	0	1	1	1	0	1
BdeBac	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	1	1	0	1	1	1	0	1
BifLon	1	1	0	1	1	2	0	2	1	2	0	2	1	1	0	3	1	1	1	1	0	1	1	1	0	1
BloFlo	1	0	0	1	1	1	0	1	1	0	0	0	1	0	0	1	1	0	1	1	0	1	1	1	0	1
BloPen	1	1	0	1	1	1	0	1	1	0	0	1	1	0	0	1	0	0	1	1	0	1	1	1	0	1
BorBur	1	0	0	1	1	0	0	0	1	0	0	0	1	0	0	1	1	0	1	1	0	1	1	0	0	1
BorGar	1	0	0	1	1	0	0	0	1	0	0	0	1	0	0	1	1	0	1	1	0	1	1	0	0	1
BorPer	1	1	0	1	1	1	0	1	1	1	0	1	3	1	0	2	1	1	1	1	0	1	1	2	0	1
BraJap	1	1	0	1	1	1	0	1	1	1	0	1	1	1	0	1	1	1	1	1	0	1	1	1	0	1
BruMel	1	1	0	1	1	1	0	1	1	1	0	1	3	1	0	2	1	1	1	1	0	2	1	1	0	1
BuchAp	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	1	0	0	1	1	0	0	1
BuchBp	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	1	0	0	1	1	0	0	1
BuchSg	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	1	0	0	0	1	0	0	1
BurPse	1	1	0	1	1	1	0	2	1	1	0	1	4	1	0	2	1	1	0	1	0	2	1	2	0	1
CamJej	1	0	0	1	1	0	0	1	2	0	0	1	3	0	0	2	1	0	1	1	0	1	1	0	0	1
CauCre	1	1	0	1	1	1	0	1	1	1	0	1	2	1	0	2	1	1	1	1	0	1	1	1	0	1
ChlAbo	1	0	0	1	1	1	0	1	1	0	0	1	1	0	0	1	1	0	1	1	0	1	1	1	0	1
ChlCav	1	0	0	1	1	1	0	1	1	0	0	1	1	0	0	1	1	0	1	1	0	1	1	1	0	1
ChlMur	1	0	0	1	1	1	0	1	1	0	0	1	1	0	0	1	1	0	1	1	0	1	1	1	0	1
ChlPne	1	0	0	1	1	1	0	1	1	0	0	1	1	0	0	1	1	0	1	1	0	1	1	1	0	1
ChlTep	1	1	0	1	1	1	0	1	1	1	0	1	2	1	0	2	1	1	2	1	0	1	1	1	0	1
ChlTra	1	0	0	1	1	1	0	1	1	0	0	1	1	0	0	1	1	0	1	1	0	1	1	1	0	1
ChrVio	1	2	0	2	1	1	0	1	5	2	0	3	8	1	0	5	1	1	1	1	0	1	1	4	0	2
CloAce	3	1	0	1	3	1	0	0	4	0	0	0	2	0	0	3	4	1	1	1	0	1	2	1	0	1
CloPer	2	0	0	1	4	0	0	0	4	0	0	0	6	0	0	4	7	0	4	1	0	1	3	0	0	1
CloTet	1	0	0	1	3	0	0	0	1	0	0	0	3	0	0	2	2	0	2	1	0	1	1	0	0	1
ColPsy	4	0	0	1	2	0	0	1	6	0	0	1	7	0	0	4	1	0	3	1	0	1	2	0	0	1
CorDip	1	1	0	1	1	1	0	2	1	1	0	1	3	0	0	3	1	1	1	1	0	1	1	1	0	1
CorEff	1	1	0	1	1	1	0	2	1	1	0	2	1	0	0	3	1	1	1	1	0	2	1	1	0	1
CorGlu	1	1	0	2	1	1	0	2	1	1	0	1	3	0	0	3	1	1	1	1	0	2	1	1	0	1
CorJej	1	1	0	1	1	1	0	2	1	1	0	1	1	0	0	3	1	1	1	1	0	1	1	1	0	1
CoxBur	1	1	0	1	1	1	0	1	1	0	0	1	1	0	0	1	1	1	1	1	0	1	1	1	0	1
DecAro	1	1	0	2	1	1	0	1	2	1	0	2	4	1	0	4	1	1	1	1	0	2	1	2	0	1
DehEth	1	1	0	1	1	1	0	1	1	1	0	1	1	1	0	1	1	1	1	1	0	1	1	1	0	1
DeiRad	1	1	0	1	1	1	0	1	1	1	0	1	1	1	0	3	1	1	1	1	0	1	1	1	0	1
DesPsy	1	0	0	2	1	1	0	1	6	0	0	1	2	0	0	2	1	0	1	1	0	2	1	1	0	1
DesVul	1	1	0	2	1	1	0	1	1	1	0	2	5	1	0	3	1	1	1	1	0	2	1	2	0	1
EhrCan	1	0	0	1	1	0	0	1	1	0	0	0	1	0	0	1	1	0	1	1	0	1	1	1	0	1
EhrRum	1	0	0	1	1	0	0	1	1	0	0	0	1	0	0	1	1	0	1	1	0	1	1	1	0	1
EntFae	2	0	0	1	2	1	0	0	3	0	0	0	3	0	0	2	2	0	2	1	0	1	2	0	0	1
ErwCar	2	0	0	2	1	1	0	2	3	0	0	2	3	0	0	4	1	1	1	1	0	1	1	4	0	2
EscCol	1	1	0	2	1	2	0	2	5	0	0	2	3	0	0	4	1	1	1	1	0	1	1	4	0	2
FraTul	1	0	0	1	1	0	0	1	2	0	0	0	3	0	0	2	1	0	1	1	0	1	1	0	0	1
FusNuc	2	0	0	0	2	0	0	0	3	0	0	0	2	0	0	1	3	0	1	1	0	0	2	0	0	0
GeoKau	2	2	0	1	2	2	0	2	3	0	0	2	5	1	0	5	2	1	2	1	0	2	1	1	0	1
GeoSul	1	1	0	1	1	1	0	1	1	1	0	1	2	1	0	1	1	1	1	1	0	1	1	1	0	1
GloVio	1	1	0	1	1	1	0	1	1	1	0	1	1	1	0	1	1	1	1	1	0	1	1	1	0	1
GluOxy	1	1	0	1	1	1	0	1	1	1	0	1	4	1	0	2	1	1	1	1	0	1	1	1	0	1

Amino Acid	Ser	Ser	Ser	Ser	Arg	Arg	Arg	Arg	Arg	Arg	Totals
AntiCodon	UGA	CGA	ACU	GCU	ACG	GCG	UCG	CCG	UCU	CCU	
AciAdp	2	0	0	1	3	0	0	1	1	1	68
AgrTum	1	1	0	1	1	0	0	1	1	0	46
AnaMar	1	0	0	1	1	0	0	1	1	1	33
AquAeo	1	1	0	1	1	0	0	1	1	1	40
AzoEbn	1	1	0	1	1	0	0	1	1	1	54
BacAnt	4	0	0	2	3	0	0	1	1	0	85
BacCla	3	0	0	2	3	0	0	1	1	1	68
BacFra	2	0	0	1	3	0	0	1	1	1	67
BacHal	3	0	0	1	4	0	0	1	1	0	71
BacLic	2	0	0	2	2	0	0	1	1	1	64
BacSub	2	0	0	2	4	0	0	1	1	1	79
BacThe	2	0	0	1	3	0	0	1	1	1	65
BarHen	1	1	0	1	1	0	0	1	1	0	37
BarQui	1	1	0	1	1	0	0	1	1	0	37
BdeBac	1	0	0	1	1	0	1	0	1	0	32
BifLon	1	1	0	1	2	0	0	1	1	1	51
BloFlo	1	1	0	1	1	0	0	1	1	1	33
BloPen	1	1	0	1	1	0	0	1	1	1	35
BorBur	1	0	0	1	0	1	1	0	1	0	28
BorGar	1	0	0	1	0	1	1	0	1	0	28
BorPer	0	1	0	1	2	0	0	1	1	1	46
BraJap	1	1	0	1	1	0	1	1	1	1	44
BruMel	1	1	0	1	1	0	0	1	1	1	48
BuchAp	1	0	0	1	1	0	0	1	1	0	27
BuchBp	1	0	0	1	1	0	0	1	1	0	27
BuchSg	1	0	0	1	1	0	0	1	1	0	26
BurPse	1	2	0	1	2	0	0	1	1	1	55
CamJej	1	0	0	1	0	1	1	0	2	1	39
CauCre	1	1	0	1	1	0	0	1	1	1	45
ChlAbo	1	1	0	1	1	0	1	0	1	1	34
CHlCav	1	1	0	1	1	0	1	0	1	1	34
ChlMur	1	1	0	1	1	0	1	0	1	0	33
ChlPne	1	1	0	1	1	0	1	0	1	1	34
ChlTep	1	1	0	1	1	0	0	1	1	1	45
ChlTra	1	1	0	1	1	0	1	0	1	0	33
ChrVio	1	1	0	1	3	0	0	1	1	1	92
CloAce	1	1	0	1	1	0	1	0	3	1	67
CloPer	2	0	0	3	1	0	1	0	3	1	86
CloTet	1	1	0	1	1	0	1	0	1	1	48
ColPsy	3	0	0	1	3	0	0	1	1	0	77
CorDip	1	1	0	1	2	0	0	1	1	1	48
CorEff	1	1	0	1	2	0	0	1	1	1	51
CorGlu	1	1	0	1	2	0	0	1	1	1	55
CorJej	1	1	0	1	2	0	0	1	1	1	46
CoxBur	1	1	0	1	1	0	1	0	1	1	38
DecAro	1	1	0	1	2	0	0	1	1	1	59
DehEth	1	1	0	1	0	1	1	1	1	1	42
DeiRad	1	1	0	1	1	0	0	1	1	1	45
DesPsy	1	0	0	1	2	0	1	0	1	1	58
DesVul	1	1	0	1	3	0	0	1	1	1	62
EhrCan	1	0	0	1	1	0	0	1	1	1	32
EhrRum	1	0	0	1	1	0	0	1	1	1	32
EntFae	2	1	0	1	2	0	0	1	1	1	55
ErwCar	2	0	0	1	3	0	0	1	1	1	68
EscCol	1	1	0	1	4	0	0	1	1	1	78
FraTul	1	0	0	1	1	0	0	1	1	0	34
FusNuc	2	0	0	1	1	0	1	0	1	1	43
GeoKau	1	2	0	2	4	0	0	1	1	1	81
GeoSul	1	2	0	1	1	0	0	1	1	1	45
GloVio	1	1	0	1	1	0	0	1	1	1	41
GluOxy	1	1	0	1	2	0	0	1	1	0	48

Appendix D

Amino Acid	Met	Trp	Phe	Phe	Tyr	Tyr	Ile	Ile	Ile	Asn	Asn	His	His	Asp	Asp	Cys	Cys	Sel	Gln	Gln	Lys	Lys	Glu	Glu	Pro	Pro
AntiCodon	CAU	CCA	AAA	GAA	AUA	GUA	AAU	GAU	UAU	AUU	GUU	AUG	GUG	AUC	GUC	ACA	GCA	UCA	UUG	CUG	UUU	CUU	UUC	CUC	AGG	GGG
HaeDuc	4	1	0	1	0	1	0	3	0	0	1	0	1	0	2	0	1	1	1	0	2	0	3	0	0	0
HaeInf	4	1	0	1	0	1	0	3	0	0	2	0	1	0	3	0	1	1	2	0	3	1	3	0	0	0
HelHep	3	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1	1	1	0	0	1
HelPyl	3	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	2	0	0	1
IdiLoi	4	2	0	1	0	2	0	4	0	0	2	0	1	0	2	0	1	0	2	0	2	0	3	0	0	1
LacAci	4	1	0	2	0	2	0	2	0	0	3	0	2	0	3	0	1	0	3	1	1	2	2	1	0	0
LacJoh	7	1	0	4	0	3	0	4	0	0	4	0	1	0	3	0	1	0	3	1	2	2	4	1	0	0
LacLac	4	1	0	2	0	1	0	2	0	0	2	0	1	0	2	0	1	0	2	0	3	1	2	0	0	0
LacPla	5	2	0	2	0	2	0	3	2	0	5	0	2	0	3	0	1	0	2	1	2	1	2	1	0	0
LegPne	3	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	0	2	1	2	0	0	1
LeiXyl	3	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1	0	1
LepInt	3	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	1
LisMon	4	1	0	2	0	2	0	3	0	0	4	0	2	0	3	0	1	0	2	0	3	1	4	0	0	0
ManSuc	4	1	0	1	0	1	0	3	0	0	2	0	1	0	3	0	1	1	2	0	3	1	3	0	0	0
MesFlo	3	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	0	2	0	1	0	0	0
MesLot	4	1	0	1	0	1	0	2	1	0	1	0	1	0	1	0	1	0	2	1	1	1	1	1	0	1
MetCap	3	1	0	1	0	1	0	2	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1	0	1
MycAvi	3	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	1	1	1	1	1	1	1	0	1
MycGal	3	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1	1	1	0	0	0
MycGen	3	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1	1	1	0	0	0
MycHyo	3	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	0
MycLep	3	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1	0	1
MycMob	2	1	0	1	0	1	0	1	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	0
MycMyc	3	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	0	2	1	1	0	0	0
MycPen	3	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1	1	1	0	0	0
MycPne	3	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1	1	1	0	0	0
MycPul	3	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	0
MycSyn	5	1	0	2	0	1	0	1	0	0	1	0	1	0	2	0	1	0	1	0	1	0	1	0	0	0
MycTub	3	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1	0	1
NeiMen	3	1	0	1	0	1	0	4	0	0	2	0	1	0	2	0	1	0	3	0	2	0	3	0	0	1
NitEur	3	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	1
NocFar	4	2	0	1	0	1	0	1	0	0	1	0	1	0	1	0	2	0	1	2	1	1	1	1	0	1
NosToc	4	1	0	2	0	2	0	3	0	0	3	0	1	0	1	0	1	0	2	1	2	1	1	0	0	1
OceIhe	5	1	0	2	0	2	0	2	0	0	4	0	2	0	3	0	1	0	3	0	4	0	3	0	0	0
Onion	3	1	0	1	0	1	0	2	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	0
ParSpp	3	1	0	1	0	1	0	0	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	1
PasMul	4	1	0	1	0	1	0	3	0	0	2	0	1	0	3	0	1	1	2	0	3	0	3	0	0	0
PelUbi	4	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	0
PhoLum	7	1	0	2	0	2	0	3	0	0	4	0	1	0	3	0	1	1	2	1	12	1	4	0	0	1
PhoPro	15	3	0	3	0	8	0	3	0	0	4	0	2	0	9	0	4	1	7	0	6	1	7	0	0	0
PorGin	3	1	0	1	0	2	0	4	0	0	2	0	1	0	2	0	1	0	1	1	1	1	2	0	0	1
Pro137	3	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	1
ProAcn	3	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1	0	1
ProMed	3	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	1
ProMit	2	1	0	1	0	1	0	2	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	1
ProNat	3	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	1
PseAer	4	1	0	1	0	1	0	4	0	0	2	0	2	0	4	0	1	1	1	0	2	0	3	0	0	1
PseFlu	5	1	0	1	0	1	0	5	0	0	2	0	1	0	4	0	1	0	2	0	2	0	6	0	0	1
PsePut	6	1	0	1	0	1	0	3	0	0	2	0	2	0	5	0	1	0	2	0	3	0	5	0	0	1
PseSyr	5	1	0	1	0	1	0	5	0	0	2	0	1	0	3	0	1	0	1	0	2	0	4	0	0	1
PsyArc	3	1	0	1	0	1	0	4	0	0	2	0	1	0	2	0	1	0	2	0	2	0	2	0	0	1
RalEut	5	1	0	1	0	1	0	4	0	0	2	0	1	0	3	0	2	0	1	0	1	2	2	0	0	1
RalSol	4	1	0	1	0	1	0	4	0	0	1	0	1	0	3	0	2	0	1	0	1	1	2	0	0	1
RhoBal	3	2	0	1	0	1	0	1	2	0	3	0	1	0	1	0	1	0	1	1	2	3	3	0	0	1
RhoPal	3	1	0	1	0	1	0	2	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1	0	1
RicCon	3	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	0
RicPro	3	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	0
RicTyp	3	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	0
SalEnt	6	1	0	2	0	3	0	3	0	0	4	0	1	0	3	0	1	1	2	2	5	0	4	0	0	1
SheOne	8	1	0	3	0	4	0	3	0	0	5	0	2	0	4	0	2	1	3	0	7	0	6	0	0	1
SilPom	4	1	0	1	0	0	0	3	0	0	2	0	1	0	2	0	1	0	2	0	1	0	2	0	0	1

Amino Acid	Pro	Pro	Thr	Thr	Thr	Thr	Val	Val	Val	Val	Ala	Ala	Ala	Ala	Gly	Gly	Gly	Gly	Leu	Leu	Leu	Leu	Leu	Leu	Ser	Ser
AntiCodon	UGG	CGG	AGU	GGU	UGU	CGU	AAC	GAC	UAC	CAC	AGC	GGC	UGC	CGC	ACC	GCC	UCC	CCC	UAA	CAA	AAG	GAG	UAG	CAG	AGA	GGA
HaeDuc	2	0	0	1	1	0	0	1	3	0	0	1	3	0	0	3	1	0	1	1	0	1	1	0	0	1
HaeInf	2	0	0	1	1	0	0	1	4	0	0	1	3	0	0	3	1	0	2	1	0	1	1	0	0	1
HelHep	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	1	1	0	1	1	0	0	1
HelPyl	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	1	1	0	1	1	0	0	1
IdiLoi	2	0	0	1	1	0	0	1	2	0	0	1	4	0	0	3	1	0	1	1	0	1	1	1	0	1
LacAci	2	1	1	1	1	1	0	0	3	0	0	0	1	0	0	3	1	1	1	1	1	0	1	1	0	1
LacJoh	2	1	2	1	2	1	0	0	3	0	0	0	2	0	0	3	2	1	2	1	1	0	2	1	0	1
LacLac	2	0	0	1	2	1	0	0	3	0	0	0	6	0	0	2	2	0	2	1	2	0	2	1	0	1
LacPla	2	1	1	1	1	1	0	0	3	0	0	0	2	1	0	2	3	1	2	1	1	0	1	1	0	1
LegPne	1	1	0	1	1	1	0	1	1	0	0	1	2	0	0	1	1	1	1	1	0	1	1	1	0	1
LeiXyl	1	1	0	1	1	1	0	1	1	1	0	1	1	1	0	1	1	1	1	1	0	1	1	1	0	1
LepInt	1	0	0	1	1	1	0	1	1	0	0	1	1	0	0	1	1	0	1	1	0	1	1	1	0	1
LisMon	2	0	0	1	2	1	0	1	3	0	0	0	4	0	0	3	2	0	2	1	0	1	2	0	0	1
ManSuc	2	0	0	1	2	0	0	1	4	0	0	1	3	0	0	4	1	0	2	1	0	1	1	1	0	1
MesFlo	1	0	1	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	1	1	0	0	1	0	0	0
MesLot	1	1	0	1	1	1	0	1	1	1	0	1	2	1	0	1	1	1	1	1	0	1	1	1	0	1
MetCap	1	1	0	1	1	1	0	1	1	1	0	1	2	0	0	1	1	1	1	1	0	1	1	1	0	1
MycAvi	1	1	0	1	1	1	0	1	1	1	0	1	1	1	0	1	1	1	1	1	0	1	1	1	0	1
MycGal	1	0	0	1	1	0	0	0	1	0	0	0	1	0	0	1	1	0	1	1	0	0	1	0	0	0
MycGen	1	0	0	1	1	1	0	0	1	0	0	0	1	0	0	1	1	0	1	1	0	1	1	0	0	1
MycHyo	1	0	0	1	1	0	0	0	1	0	0	0	1	0	0	0	1	0	1	1	0	1	1	0	0	1
MycLep	1	1	0	1	1	1	0	1	1	1	0	1	1	1	0	1	1	1	1	1	0	1	1	1	0	1
MycMob	1	0	0	1	1	0	0	0	1	0	0	0	1	0	0	0	1	0	1	1	0	0	1	0	0	0
MycMyc	1	0	1	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	1	1	0	0	1	0	0	0
MycPen	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	1	1	0	1	1	0	0	1	0	0	0
MycPne	1	0	0	1	1	1	0	0	1	0	0	0	1	0	0	1	1	0	1	1	0	1	1	0	0	1
MycPul	1	0	0	1	1	0	0	0	1	0	0	0	1	0	0	0	1	0	1	1	0	0	1	0	0	0
MycSyn	1	0	0	1	1	0	0	0	1	0	0	0	1	0	0	0	1	0	1	1	1	0	1	0	0	0
MycTub	1	1	0	1	1	1	0	1	1	1	0	1	1	1	0	1	1	1	1	1	0	1	1	1	0	1
NeiMen	1	1	0	1	1	1	0	1	2	0	0	1	4	0	0	4	1	0	1	1	0	1	1	2	0	1
NitEur	1	1	0	1	1	1	0	1	1	1	0	1	1	0	0	1	1	1	1	1	0	1	1	1	0	1
NocFar	2	1	0	1	1	1	0	1	1	1	0	1	1	1	0	2	1	1	1	1	0	2	1	1	0	1
NosToc	2	1	0	1	2	1	0	1	2	0	0	1	5	1	0	1	2	1	1	2	0	2	2	1	0	1
OceIhe	3	0	0	1	3	0	0	1	4	0	0	1	3	0	0	3	3	0	2	1	0	1	2	0	0	1
Onion	1	0	0	1	1	0	0	0	2	0	0	0	1	0	0	1	1	0	1	1	0	1	1	0	0	1
ParSpp	1	0	0	1	1	0	0	1	1	0	0	1	0	0	0	2	1	0	1	1	0	1	1	1	0	1
PasMul	2	0	0	1	1	0	0	1	4	0	0	1	3	0	0	4	1	0	2	1	0	1	1	0	0	1
PelUbi	1	0	0	1	1	0	0	1	1	0	0	0	1	0	0	1	1	0	1	1	0	0	1	0	0	1
PhoLum	2	0	0	2	2	0	0	2	3	0	0	2	3	0	0	4	1	1	1	1	0	1	1	3	0	1
PhoPro	6	0	0	2	7	0	0	4	7	0	0	2	5	0	0	11	4	0	5	1	0	1	12	0	0	2
PorGin	1	1	0	1	1	1	0	0	2	0	0	1	4	0	0	2	1	1	1	1	0	1	1	1	0	1
Pro137	1	0	0	1	1	1	0	1	1	0	0	1	1	0	0	1	1	1	1	1	1	0	1	1	0	1
ProAcn	1	1	0	1	1	1	0	1	1	1	0	1	1	1	0	1	1	1	1	1	0	1	1	1	0	1
ProMed	1	0	0	1	1	1	0	1	1	0	0	1	1	0	0	1	1	0	1	1	1	0	1	0	0	1
ProMit	1	1	0	1	1	1	0	1	1	1	0	1	2	1	0	1	1	1	1	1	0	1	1	1	0	1
ProNat	1	0	0	1	1	1	0	1	1	0	0	1	1	0	0	1	1	1	1	1	1	0	1	0	0	1
PseAer	1	1	0	1	1	1	0	1	2	0	0	2	4	0	0	3	1	1	1	1	0	1	1	2	0	1
PseFlu	2	0	0	1	1	1	0	1	3	0	0	3	5	0	0	4	1	1	1	1	0	1	1	2	0	1
PsePut	2	1	0	1	1	1	0	2	3	0	0	3	3	0	0	4	1	1	1	1	0	1	1	2	0	1
PseSyr	2	0	0	1	1	0	0	2	2	0	0	2	5	0	0	3	1	1	1	1	0	1	1	2	0	1
PsyArc	1	0	0	1	1	0	0	1	2	0	0	1	4	0	0	2	1	0	1	1	0	1	1	0	0	1
RalEut	1	1	0	2	1	1	0	2	1	1	0	2	4	1	0	3	1	1	1	1	0	1	1	3	0	1
RalSol	1	1	0	1	1	1	0	1	1	1	0	2	4	1	0	2	1	1	1	1	0	1	1	2	0	1
RhoBal	3	1	0	1	1	1	0	2	2	1	0	1	3	1	0	3	2	0	2	2	0	2	3	2	0	1
RhoPal	1	1	0	1	1	1	0	1	1	1	0	1	2	1	0	1	1	1	1	1	0	1	1	1	0	1
RicCon	1	0	0	1	1	1	0	1	1	0	0	1	1	0	0	1	1	0	1	0	0	1	1	0	0	1
RicPro	1	0	0	1	1	1	0	0	1	0	0	1	1	0	0	1	1	0	1	0	0	1	1	0	0	1
RicTyp	1	0	0	1	1	1	0	1	1	0	0	1	1	0	0	1	1	0	1	0	0	1	1	0	0	1
SalEnt	1	1	0	2	1	1	0	2	4	0	0	2	3	0	0	4	1	1	1	1	0	1	1	4	0	2
SheOne	3	0	0	2	3	0	0	2	5	0	0	2	3	0	0	6	1	0	3	1	0	1	2	2	0	1
SilPom	2	1	0	1	1	1	0	1	1	1	0	1	3	0	0	3	1	1	1	1	0	1	1	1	0	1

Amino Acid	Ser	Ser	Ser	Ser	Arg	Arg	Arg	Arg	Arg	Arg	Totals
AntiCodon	UGA	CGA	ACU	GCU	ACG	GCG	UCG	CCG	UCU	CCU	
HaeDuc	1	0	0	1	1	0	0	1	1	0	43
HaeInf	2	0	0	1	2	0	0	1	1	0	52
HelHep	1	0	0	1	0	1	1	0	1	1	32
HelPyl	1	0	0	1	0	1	1	0	1	1	32
IdiLoi	1	0	0	1	2	0	0	1	1	1	50
LacAci	2	1	0	1	2	0	0	1	1	1	56
LacJoh	2	1	0	2	2	0	0	0	1	1	70
LacLac	2	1	0	1	2	0	0	1	1	1	56
LacPla	1	1	0	1	2	0	0	1	1	1	63
LegPne	1	0	0	1	1	0	1	0	1	1	39
LeiXyl	1	1	0	1	1	0	0	1	1	1	41
LepInt	1	0	0	1	1	0	0	1	1	1	33
LisMon	1	1	0	2	2	0	0	1	1	1	62
ManSuc	2	0	0	1	2	0	0	1	1	0	55
MesFlo	1	0	0	1	1	0	0	0	1	0	24
MesLot	1	1	0	1	1	0	0	1	2	1	46
MetCap	1	1	0	1	1	0	0	1	1	1	42
MycAvi	1	1	0	1	1	0	0	1	1	1	42
MycGal	1	0	0	1	0	2	1	0	1	0	27
MycGen	1	1	0	1	0	1	1	0	1	1	31
MycHyo	1	0	0	1	1	0	0	0	1	0	25
MycLep	1	1	0	1	1	0	0	1	1	1	41
MycMob	1	0	0	1	1	0	0	0	1	0	24
MycMyc	1	0	0	1	1	0	0	0	1	0	25
MycPen	1	0	0	1	1	0	1	0	1	0	25
MycPne	1	2	0	1	0	1	1	0	1	1	32
MycPul	1	0	0	1	0	1	1	0	1	0	24
MycSyn	2	0	0	1	1	0	0	0	1	0	27
MycTub	1	1	0	1	1	0	0	1	1	1	41
NeiMen	1	1	0	1	2	0	0	1	1	1	54
NitEur	1	1	0	1	1	0	0	1	1	1	37
NocFar	1	1	0	1	1	0	0	1	1	2	47
NosToc	1	2	0	1	2	0	0	1	2	1	61
OceIhe	1	0	0	1	3	0	0	1	1	1	63
Onion	1	0	0	1	1	0	0	0	1	0	28
ParSpp	1	0	0	1	1	0	1	0	1	1	31
PasMul	2	0	0	1	2	0	0	1	1	0	52
PelUbi	1	0	0	1	1	0	1	0	1	0	27
PhoLum	2	0	0	1	3	0	0	1	1	1	77
PhoPro	10	0	0	3	10	0	0	1	1	1	150
PorGin	1	0	0	1	2	0	0	1	1	1	49
Pro137	1	1	0	1	1	0	0	1	1	1	35
ProAcn	1	1	0	1	1	0	0	1	1	1	41
ProMed	1	1	0	1	1	0	0	1	1	1	33
ProMit	1	1	0	1	1	0	0	1	1	1	40
ProNat	1	1	0	1	1	0	0	1	1	1	34
PseAer	1	1	0	1	3	0	0	1	1	1	58
PseFlu	1	1	0	1	2	0	0	1	1	1	64
PsePut	1	1	0	1	4	0	0	1	1	2	67
PseSyr	1	1	0	1	2	0	0	2	1	1	58
PsyArc	1	0	0	1	1	0	0	1	1	1	44
RalEut	1	1	0	1	2	0	0	1	1	1	59
RalSol	1	1	0	1	1	0	0	1	1	1	52
RhoBal	1	1	0	1	1	0	1	1	2	1	65
RhoPal	1	1	0	1	1	0	1	1	1	1	44
RicCon	1	0	0	1	1	0	0	1	1	0	29
RicPro	1	0	0	1	1	0	0	1	1	0	28
RicTyp	1	0	0	1	1	0	0	1	1	0	29
SalEnt	1	1	0	1	4	0	0	1	3	2	78
SheOne	3	0	0	2	6	0	0	1	1	1	92
SilPom	1	1	0	1	1	0	0	1	1	1	47

Appendix D

Amino Acid	Met	Trp	Phe	Phe	Tyr	Tyr	Ile	Ile	Ile	Asn	Asn	His	His	Asp	Asp	Cys	Cys	Sel	Gln	Gln	Lys	Lys	Glu	Glu	Pro	Pro
AntiCodon	CAU	CCA	AAA	GAA	AUA	GUA	AAU	GAU	UAU	AUU	GUU	AUG	GUG	AUC	GUC	ACA	GCA	UCA	UUG	CUG	UUU	CUU	UUC	CUC	AGG	GGG
SinMel	6	1	0	1	0	1	0	3	0	0	1	0	1	0	2	0	1	1	1	1	1	1	3	1	0	1
StaAur	4	1	0	2	0	2	0	3	0	0	3	0	2	0	4	0	1	0	2	0	3	0	3	0	0	0
StaEpi	4	1	0	2	0	2	0	2	0	0	3	0	2	0	3	0	1	0	2	0	3	0	3	0	0	0
StaHae	4	1	0	2	0	2	0	2	1	0	3	0	2	0	4	0	1	0	2	0	3	0	3	0	0	0
StaSap	4	1	0	2	0	2	0	2	0	0	3	0	2	0	3	0	1	0	2	0	3	0	3	0	0	0
StrAga	5	2	0	3	0	3	0	4	0	0	3	0	2	0	3	0	1	0	4	0	3	1	5	0	0	0
StrAve	6	1	0	1	0	1	0	1	0	0	2	0	1	0	2	0	2	0	1	2	1	3	1	3	0	3
StrCoe	5	1	0	1	0	1	0	1	0	0	2	0	1	0	2	0	1	0	0	2	1	3	1	3	0	3
StrPne	4	1	0	2	0	2	0	2	0	0	2	0	1	0	2	0	1	0	2	0	2	1	5	0	0	0
StrPyo	2	1	0	2	0	2	0	4	0	0	3	0	1	0	2	0	1	0	2	0	2	1	4	0	0	0
StrThe	4	1	0	2	0	2	0	2	0	0	3	0	1	0	3	0	1	0	3	0	3	1	4	0	0	0
SymThe	6	2	0	3	0	3	0	4	0	0	3	0	2	0	4	0	2	1	1	2	1	4	2	3	0	2
Syn680	2	1	0	1	0	1	0	2	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	1
SynElo	3	1	0	1	0	1	0	2	0	0	1	0	1	0	1	0	2	0	1	0	1	0	1	0	0	1
SynSpp	2	1	0	1	0	1	0	2	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	1
TheElo	2	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	1
TheFus	4	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	1	1	2	1	2	0	1
TheMar	3	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1	0	1
TheTen	4	1	0	2	0	2	0	1	0	0	3	0	1	0	2	0	1	0	1	1	2	1	2	1	0	1
TheThe	3	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	2	0	1
TreDen	3	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	1	1	1	1	1	1	1	0	1
TrePal	3	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1	0	1
TroWhi	3	1	0	3	0	1	0	1	0	0	1	0	1	0	2	0	1	0	1	1	1	1	3	1	0	1
UreUre	3	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1	1	1	0	0	0
VibCho	9	1	0	3	0	5	0	3	0	0	4	0	2	0	5	0	3	0	5	0	2	0	4	0	0	1
VibFis	8	3	0	4	0	4	0	3	0	0	7	0	2	0	6	0	1	0	5	0	4	0	6	0	0	0
VibPar	11	2	0	4	0	7	0	2	0	0	5	0	2	0	6	0	4	0	6	0	4	0	6	0	0	0
VibVul	9	2	0	4	0	4	0	3	0	0	5	0	2	0	6	0	3	0	4	0	5	0	5	0	0	1
wigGlo	3	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	2	0	0	0
WolPip	3	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	0
WolSuc	3	1	0	1	0	1	0	3	0	0	1	0	1	0	1	0	1	0	1	0	1	0	2	0	0	1
WolTRS	3	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	0
XanAxo	3	1	0	1	0	1	0	2	0	0	1	0	1	0	2	0	1	0	1	0	1	1	2	1	0	1
XanCam	3	1	0	1	0	1	0	2	0	0	1	0	1	0	2	0	1	0	1	1	1	1	2	1	0	1
XanOry	2	1	0	1	0	1	0	1	0	0	1	0	1	0	2	0	1	0	1	1	1	1	0	1	0	1
XylFas	3	1	0	1	0	1	0	2	0	0	1	0	1	0	1	0	1	0	1	1	2	1	1	1	0	1
YerPes	6	1	0	2	0	2	0	3	0	0	3	0	1	0	3	0	1	1	1	1	3	1	3	0	0	1
ZymMob	5	1	0	1	0	1	0	0	0	0	1	0	1	0	1	0	1	0	2	1	0	1	0	1	0	0

Amino Acid	Pro	Pro	Thr	Thr	Thr	Thr	Val	Val	Val	Val	Ala	Ala	Ala	Ala	Gly	Gly	Gly	Gly	Leu	Leu	Leu	Leu	Leu	Leu	Ser	Ser
AntiCodon	UGG	CGG	AGU	GGU	UGU	CGU	AAC	GAC	UAC	CAC	AGC	GGC	UGC	CGC	ACC	GCC	UCC	CCC	UAA	CAA	AAG	GAG	UAG	CAG	AGA	GGA
SinMel	1	1	0	1	1	1	0	1	1	0	0	1	3	1	0	1	1	1	1	1	0	1	1	1	0	1
StaAur	2	0	0	0	3	0	0	0	3	0	0	0	4	0	0	2	3	0	2	1	0	1	1	0	0	1
StaEpi	2	0	0	0	3	0	0	0	3	0	0	0	3	0	0	2	3	0	2	1	0	1	1	0	0	1
StaHae	2	0	0	0	3	0	0	0	3	0	0	0	3	0	0	2	3	0	2	1	0	1	1	0	0	1
StaSap	2	0	0	0	3	0	0	0	3	0	0	0	3	0	0	2	3	0	2	1	0	1	1	0	0	1
StrAga	2	0	0	1	3	0	0	0	5	0	0	0	7	0	0	2	3	0	2	2	1	0	3	0	0	1
StrAve	1	1	0	2	1	1	0	3	1	3	0	2	1	1	0	3	1	1	1	1	0	2	1	1	0	1
StrCoe	1	1	0	2	1	1	0	3	1	2	0	2	1	1	0	3	1	1	1	1	0	2	1	1	0	1
StrPne	2	0	0	1	2	0	0	0	3	0	0	0	4	0	0	2	2	0	2	1	1	0	2	0	0	1
StrPyo	1	0	0	1	2	0	0	0	4	0	0	0	6	0	0	1	3	0	1	1	1	0	2	0	0	1
StrThe	2	0	0	1	3	0	0	0	4	0	0	0	6	0	0	2	2	0	2	1	1	0	3	1	0	1
SymThe	1	2	0	3	1	2	0	3	1	3	0	4	1	3	0	4	1	2	1	1	0	2	1	4	0	2
Syn680	1	1	0	1	1	1	0	1	1	0	0	1	1	1	0	1	1	1	1	1	0	1	1	1	0	1
SynElo	1	1	0	1	1	1	0	1	1	1	0	1	2	1	0	1	1	1	0	1	0	1	1	1	0	1
SynSpp	1	1	0	1	1	1	0	1	1	1	0	1	2	1	0	1	1	1	1	1	0	1	1	1	0	1
TheElo	1	1	0	1	1	1	0	1	1	1	0	1	1	1	0	1	1	1	0	1	0	1	1	1	0	1
TheFus	1	1	0	1	1	1	0	2	1	1	0	2	1	1	0	3	1	1	1	1	0	1	1	1	0	1
TheMar	1	1	0	1	1	1	0	1	1	1	0	1	1	1	0	1	1	1	1	1	0	1	1	1	0	1
TheTen	1	1	0	1	1	1	0	1	2	1	0	1	2	0	0	2	1	1	1	1	0	1	1	1	0	1
TheThe	1	1	0	1	1	1	0	1	1	1	0	1	1	1	0	2	1	1	1	1	0	1	1	1	0	1
TreDen	1	1	0	1	1	1	0	1	1	0	0	1	1	1	0	1	1	0	1	1	0	1	1	1	0	1
TrePal	1	1	0	1	1	1	0	1	1	1	0	1	1	1	0	1	1	0	1	1	0	1	1	1	0	1
TroWhi	1	1	0	1	1	1	0	1	1	1	0	1	1	1	0	1	1	1	1	1	0	1	1	1	0	1
UreUre	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	1	1	0	1	1	0	0	1	0	0	0
VibCho	3	0	0	2	4	0	0	2	2	0	0	1	4	0	0	6	2	0	2	1	0	1	5	3	0	1
VibFis	3	0	0	2	6	0	0	0	4	0	0	1	4	0	0	7	2	0	1	1	0	1	6	0	0	1
VibPar	3	0	0	2	5	0	0	2	4	0	0	1	4	0	0	11	2	0	3	1	0	2	10	0	0	1
VibVul	3	0	0	2	4	0	0	2	4	0	0	1	5	0	0	7	2	0	2	1	0	2	7	0	0	1
wigGlo	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	1	1	0	1	1	0	0	1
WolPip	1	0	0	1	1	0	0	1	1	0	0	0	1	0	0	1	1	0	1	1	0	1	1	1	0	1
WolSuc	1	0	0	1	1	0	0	1	1	0	0	1	3	0	0	1	1	0	1	1	0	1	1	0	0	1
WolTRS	1	0	0	1	1	0	0	1	1	0	0	0	1	0	0	1	1	0	1	1	0	1	1	1	0	1
XanAxo	1	1	0	1	1	1	0	1	1	1	0	2	3	1	0	2	1	1	1	1	0	1	1	2	0	1
XanCam	1	1	0	1	1	1	0	1	1	1	0	2	2	1	0	2	1	1	1	1	0	1	1	2	0	1
XanOry	1	0	0	1	0	1	0	1	1	0	0	0	1	0	0	2	1	1	1	1	0	1	1	2	0	1
XylFas	1	1	0	1	1	1	0	1	1	1	0	1	2	1	0	2	1	1	1	1	0	1	1	1	0	1
YerPes	2	0	0	2	2	0	0	2	3	0	0	2	3	0	0	2	1	1	1	1	0	1	1	2	0	2
ZymMob	0	1	0	1	0	1	0	1	1	1	0	0	0	0	0	0	1	1	0	1	0	1	1	1	0	1

Amino Acid	Ser	Ser	Ser	Ser	Arg	Arg	Arg	Arg	Arg	Arg	Totals
AntiCodon	UGA	CGA	ACU	GCU	ACG	GCG	UCG	CCG	UCU	CCU	
SinMel	1	1	0	1	1	0	0	1	1	1	48
StaAur	3	0	0	1	2	0	0	1	1	0	56
StaEpi	3	0	0	1	2	0	0	1	1	0	53
StaHae	3	0	0	1	2	0	0	1	1	0	55
StaSap	3	0	0	1	2	0	0	1	1	0	53
StrAga	3	0	0	1	2	0	0	1	1	1	73
StrAve	1	1	0	2	1	0	0	1	1	1	61
StrCoe	1	1	0	1	1	0	0	1	1	1	57
StrPne	2	0	0	1	2	0	0	1	1	1	53
StrPyo	2	0	0	1	3	0	0	1	1	1	57
StrThe	2	0	0	1	2	0	0	1	1	1	62
SymThe	1	2	0	2	2	0	0	2	1	1	90
Syn680	1	1	0	1	1	0	0	1	1	1	38
SynElo	1	1	0	1	1	0	0	1	1	1	40
SynSpp	1	1	0	1	1	0	0	1	1	1	40
TheElo	1	1	0	1	1	0	0	1	1	1	37
TheFus	1	1	0	1	1	0	0	1	1	1	47
TheMar	1	1	0	1	0	1	1	1	1	1	42
TheTen	1	1	0	1	1	0	0	1	1	1	50
TheThe	1	1	0	1	1	0	0	1	1	1	43
TreDen	1	0	0	0	0	1	1	0	1	1	38
TrePal	1	1	0	1	0	1	1	1	1	1	41
TroWhi	1	1	0	1	1	0	0	1	1	1	46
UreUre	1	0	0	1	1	0	1	0	1	0	25
VibCho	2	0	0	2	6	0	0	1	1	0	88
VibFis	2	0	0	2	7	0	0	1	0	0	93
VibPar	4	0	0	1	8	0	0	1	1	1	113
VibVul	4	0	0	2	6	0	0	1	1	1	100
wigGlo	1	0	0	1	1	0	0	1	1	0	30
WolPip	1	0	0	1	1	0	0	1	1	1	30
WolSuc	1	0	0	1	0	1	1	0	1	1	36
WolTRS	1	0	0	1	1	0	0	1	1	1	30
XanAxo	1	1	0	1	2	0	0	1	1	1	49
XanCam	1	1	0	1	2	0	0	1	1	1	49
XanOry	1	1	0	1	2	0	0	0	1	1	38
XylFas	1	1	0	1	1	0	0	1	1	1	45
YerPes	2	0	0	1	2	0	0	1	1	1	62
ZymMob	1	0	0	1	0	0	0	1	1	1	29

Appendix D