Analysis of Codon Usage

by

John F. Peden B.Sc., M.Sc.

Department of Genetics

Thesis submitted to the University of Nottingham for the Degree of Doctor of Philosophy, August, 1999

Table of Contents

Table of Tables

Table of Figures

ABSTRACT

Synonymous codons are not used randomly and it has been previously shown that in many prokaryotes and some lower eukaryotes, that in addition to mutational biases, natural selection is also influencing this non-random codon usage. This selection is for a subset of “optimal” codons in those genes that are more highly expressed. The *in-silico* analysis of codon usage has previously been hampered by a lack of suitable software. This study reports the development and application of a portable software package – CodonW – a package written in ANSI C that was specifically designed to analyse codon and amino acid usage. CodonW includes routines for correspondence analysis, the most commonly used multivariate analysis method for the analysis of codon and amino acid usage. CodonW was then applied to the analysis of codon and amino acid usage, to answer a wide range of novel biological questions.

One of the initial applications of CodonW was to re-examine the codon usage of *Saccharomyces cerevisiae.* This analysis demonstrated that between 3-4% of *S. cerevisiae* genes were highly expressed and that the distribution of codon usage indices was unimodal, skewed but was not bimodal. This study also demonstrated that it is possible to rapidly identify optimal codons and highly expressed genes with a surprisingly high degree of accuracy when analysing a whole-genome gene complement. However, it was demonstrated that care must be taken when deciding the number of genes to include as a representative group of putatively highly expressed genes, in *S. cerevisiae* it was demonstrated that this was approximately 50 genes.

This study reports the first identification of optimum codons from a cyanobacterium. Optimal codons were identified for *Synechococcus sp.* and *Synechocystis* PCC6803. The codon usage of both species is broadly similar but there is a difference in the choice of optimal codons for Gln and Arg. Genes involved in photosynthesis have a high frequency of optimal codons and appear to be adapted for translational efficiency.

The amino acid usage of *Salmonella typhimurium* was investigated using CodonW for the presence of a correlation between amino acid usage and expression levels, a correlation that has been previously reported for its close relative *Escherichia coli*. This investigation confirmed, as for *E. coli*, that the most important trend in the variation of *S. typhimurium* amino acid usage is for the usage of hydrophobic amino acids, and the third most important trend is for aromaticity. However, in *S. typhimurium,* the correlation between expression level and amino acid usage (the second most important trend in *E. coli)* is much weaker than previously reported for *E. coli* and it is uncertain if translational selection is choosing between amino acids in this species.

This study also reports the first identification of optimal codons in the Gram-positive species *Lactococcus lactis* (17 optimal codons), *Streptococcus mutans* (12 optimal codons), and *Staphylococcus aureus* (15 optimal codons). The choices of optimal codons were broadly similar with the exception of Leu. In general the codon usage of genes from *L. lactis* lactic acid fermentation pathways show a higher frequency of optimal codons relative to their *B. subtilis* homologues.

The PEPTS-*lac* operon has to date only been reported in *L. lactis*, *S. aureus* and *S. mutans*. The codon usage of these operons is similar to that of its host implying that the operons may have been present in their current hosts for a considerable period of time. There is a higher frequency of optimal codons in the *L. lactis* and *S. aureus* lac operon genes *lacD*, *lacC*, *lacB* and *lacA*. Correspondence analysis confirms that the codon usage of these genes is similar to the codon usage of highly expressed genes in these species. The codon usage of the *lac* operon appears to be under selection for translational efficiency. The variation in codon usage between the genes of the *L. lactis* and *S. aureus lac* operons strongly suggests that they have a range of expression levels. The codon usage of the *S. mutans lac* operon does not appear to be as adapted for translational efficiency.

Acknowledgements

I would like to thank all my friends who have supported me through my write-up, without whom I would never have been able to complete this thesis. In particular I extend my thanks to my supervisor, Paul, who had the generosity and foresight to offer me a placement in his laboratory, except for this I would probably be in a lab slaving over a hot gel instead of here. I would thank the genetics department in Trinity College Dublin, where I started this research project and where I shared my bench space with some interesting dental specimens. Many thanks to Andrew Lloyd, a spot of sanity in an insane world (or should that be the other way round?) - may your goats always produce a rich milk. To David and Al with whom I shared many a long evening ducking and diving, whilst slaving over a hot keyboard, all I can say is BFG. I could not forget Liz, a woman who got me into an expensive car habit.

There are many people on Oxford who have been very kind to me since my arrival, but Sarah and Bridget deserve special mention as they had to proof read as well as goad me along, many thanks. Then there is Nigel, one of the most modest people I have ever met but then he is a medic. I can’t neglect to mention Geoff, James and Michele who helped spur me on over the last hurdles.

I dedicate this thesis to those who believed in me, my parents Frank and Vivien Peden and to Sarah.

1.1 Introduction

The genetic code uses 64 codons to represent the 20 standard amino acids and the translation termination signal. Each codon is recognised by a subset of a cell’s transfer ribonucleotide acid molecules (tRNAs) and with the exception of a few codons that have been reassigned in some lineages (Osawa and Jukes 1989; Osawa *et al*. 1990) the genetic code is remarkably conserved, although it is still in a state of evolution (Osawa *et al*. 1992).

In general, codons can be grouped into 20 disjoint families, one family for each of the standard amino acids, with a 21st family for the translation termination signal. Each family in the universal genetic code contains between 1 and 6 codons. Where present, alternate codons are termed as synonymous. Although choice among synonymous codons might not be expected to alter the primary structure of a protein, it has been known for the past 20 years that alternative synonymous codons are not used randomly. This in itself is not startling as codon usage might be expected to be influenced at the very least, by mutational biases (Sharp and Matassi 1994).

The hypothesis that natural selection might be able to select between synonymous codons (also known as synonyms) is not new. Ames and Hartmann (1963) proposed that the use of alternative synonyms might have a role in the regulation of gene expression. The proposal of the neutral theory of molecular evolution by Kimura (1968) started an intense debate amongst evolutionary biologists. To test this theory there was considerable interest in the identification of sites that were not subject to Darwinian selection. King and Jukes (1969) suggested that in the absence of mutational bias synonymous codons might be used randomly, this implied that synonymous mutations be evolving neutrally. However their basic proposition, that there was no selective difference between synonyms, was strongly challenged by Clarke (1970), who advanced several mechanisms whereby Darwinian selection could choose between synonymous codons.

The first gene sequences, albeit partial, were published in the early 1970s (for a review see Sanger *et al.* 1977). As the volume of sequence data began to increase, it was suggested that in some vertebrate and invertebrate tissues, protein amino acid frequencies and tRNA concentrations were co-adapted (Chavancy *et al*. 1979; Kafatos *et al*. 1977; Suzuki and Brown 1972). This adaptation apparently varied across a wide range of cell types and concomitantly with amino acid composition and with the subcellular location of translation (Garel 1974; Maenpaa and Bernfield 1975). It was suggested that tRNA availability might regulate haemoglobin synthesis in developing blood cells (Smith 1975). Differences in the substitution rates between the conserved and variant segments of beta-globin were attributed to differences in selective constraints of mRNA secondary structure (Kafatos, Efstratiadis and Forget 1977). A negative correlation was found between mRNA stability (half-life) and frequency of rare codons, it was presumed that selection for stable mRNAs was either the same, or acted in parallel with, selection for the avoidance of non-optimal codons (Herrick *et al*. 1980). While a correlation between amino acid usage and tRNA frequencies appeared to be adaptive, in multicellular eukaryotes it is the exception rather than the norm, and is restricted to a relatively small number of proteins and cell types (Chavancy and Garel 1981).

Analysis of genes from the RNA bacteriophage MS2 identified differences between the codon usage of phage genes and genes from its host, *E. coli* (Elton *et al*. 1976; Fiers *et al*. 1975). Fiers *et al.* (1975) suggested that the observed codon bias in MS2 might result from selection for the rate of chain elongation during protein translation (Fiers *et al*. 1976; Fiers *et al*. 1975). Fitch (1976) noted a significant bias for cytosine (C) over uracil (U), and suggested that there may be selection against codon wobble pairing, avoidance of wobble pairing was also noted in yeast (Bennetzen and Hall 1982). It was suggested that the most frequent synonyms of MS2 were those translated by the major tRNAs of its host (Elton, Russell and Subak-Sharpe 1976). The observation of codon usage bias implied that not all synonymous mutations were neutral (Berger 1977). The codon usage of the bacteriophage ΦX174 (5,386 bp), the first genome to be sequenced entirely (Sanger *et al*. 1977), was found to be non-random, with a bias towards codons whose third position was thymidine (T) and away from codons starting with adenosine (A) or guanidine (G) (Sanger *et al*. 1977).

Pedersen *et al*. (1978) suggested that *Escherichia coli* codons might be translated at different rates. Post *et al*. (1979) noted that in *E. coli* there was a stronger bias in codon usage in the highly expressed ribosomal protein genes than in the weakly expressed regulatory gene *lacI*. It was also noted that the preferred synonyms in the ribosomal protein genes were recognised by abundant tRNA species and it was suggested this may be the result of selection for fidelity (Post *et al*. 1979). The constraint of maintaining a stable RNA secondary structure was suggested as another influence on codon bias (Hasegawa *et al*. 1979). A strong correlation between GC_{3s} (G+C content at the third position of synonymous codons) and the genomic G+C composition in the *trpG* gene region of a number of enterobacterial species indicated that, the choice of synonymous codons was, at least in part, influenced by the same factors that caused genomic G+C content to differ (Nichols *et al*. 1980). The suggestion that codons that have the potential to mutate to termination codons in a single step would be avoided (Modiano *et al*. 1981) has been rejected because the selective advantage of such a strategy, if it existed, would be too small to significantly influence codon usage and would involve second generation selection (Kimura 1983).

The genetic sequence databases such as EMBL (Emmert *et al*. 1994), GenBank (Benson *et al*. 1994), PIR (George *et al*. 1994), and SwissProt (Bairoch and Boeckmann 1994) have become an invaluable source of sequence information. While many of the early sequences were submitted as discrete gene fragments, genes or operons this is being rapidly superseded by the submission of entire chromosomes, genomes, and proteomes *en bloc* from dedicated genome projects. These projects have resulted in a significant increase in the quality of sequence data available.

During the early 1990's it was generally thought that the first genomes of free-living organisms to be sequenced would be *E. coli* and *S. cerevisiae.* However, the first free-living organism to be completely sequenced was *H. influenzae* (1.8 Mb) by the non-profit making TIGR corporation (Fleischmann *et al*. 1995). This was rapidly followed by the sequencing of the Gram-positive *Mycoplasma genitalium,* which possibly has the smallest genome of any free-living organism (Fraser *et al*. 1995). These demonstrations, that large scale shotgun genome sequencing projects were both feasible and cost effective, have stimulated an ever-increasing procession of genome sequencing projects. The next completed genomes were the methanogenic archaeon *Methanococcus jannaschii* (Bult *et al*. 1996), the unicellular cyanobacterium *Synechocystis* (Kaneko *et al*. 1996), and obligate parasite *Mycoplasma pneumoniae* (Himmelreich *et al*. 1996). The first eukaryotic chromosome to be sequenced was *Saccharomyces cerevisiae* chromosome III (Oliver *et al*. 1992), the sequencing of the 15 remaining chromosomes was completed by April 1996 (Goffeau *et al*. 1997). Two separate strains of the pathogen *Helicobacter pylori* have been independently sequenced (Alm *et al*. 1999; Tomb *et al*. 1997). The genomes of the model organisms *E. coli* and *Bacillus subtilis* have also been completed their progress lagged behind some of the other projects due to their greater emphasis on classical genetic mapping (Blattner *et al*. 1997; Kunst *et al*. 1997). Other completed genomes include: *Methanobacterium thermoautotrophicum* (Smith *et al*. 1997); *Archaeoglobus fulgidus* (Klenk *et al*. 1997); the spirochaetes *Borrelia burgdorferi* and *Treponema pallidum* (Fraser *et al*. 1998; Fraser *et al*. 1997); *Aquifex aeolicus* (Deckert *et al*. 1998); Mycobacterium tuberculosis (Cole *et al*. 1998) and *Rickettsia prowazekii* (Andersson *et al*. 1998).

There are more than 40 genome projects in progress including: the largest and most ambitious sequencing project “The Human Genome Project” due to have a one pass coverage completed by spring 2000 (Marshall 1999; Wadman 1999); the puffer fish *Fugu rubripes* (Aparicio *et al*. 1995); the fruit fly *Drosophila melanogaster (Rubin 1998);* the human malaria parasite *Plasmodium falciparum* (Gardner *et al*. 1998); the model plant *Arabidopsis thaliana* (Bevan *et al*. 1999); and mouse (Blake *et al*. 1999). Every genome has a unique story to tell and will advance the understanding of genome evolution, genome comparison will help to resolve many questions about genome evolution.

Perhaps one of the most surprising results is that so many of the genes that have been identified as putatively encoding protein (partially based on codon usage) have no known function or homologue. Between 15 and 20 percent of the potentially coding open reading frame (ORFs) remain unidentified and have no detectable sequence identity with another protein. This is perhaps most surprising in the case of *M. genitalium* as it was sequenced because it has the smallest genome known to be self- replicating and presumably is, or has been, under selection to minimise its gene compliment (Bloom 1995).

1.2 Natural Selection of Codon Usage

The exponential increase in the volume of sequence information during the early 1980s facilitated for the first time detailed statistical analyses of codon usage. Multivariate analysis techniques were applied to the analysis of the codon usage in mammalian, viral, bacteriophage, bacterial, mitochondrial and lower eukaryote genes (Grantham *et al*. 1980a; Grantham *et al*. 1981; Grantham *et al*. 1980b). The results of Grantham and co-workers demonstrated that genes could be grouped based on their codon usage and that these groups agreed broadly with taxonomic groupings. Consequently, they proposed the Genome Theory, which was "that the codon usage pattern of a genome was a specific characteristic of an organism". Compilations of codon usage information have confirmed broadly this organism specific codon choice pattern (Aota *et al*. 1988; Aota and Ikemura 1986). It was suggested that this variation in codon usage might be correlated with variation in tRNA abundance (Grantham *et al*. 1980b), and that this might "modulate" gene expression (Grantham *et al*. 1981).

The non-random usage of codons and variation in codon usage between species suggested some selective constraint on codon choice. The codon usage of thirteen strongly and sixteen weakly expressed *E. coli* genes was examined, again using a multivariate analysis technique, and was found to have a marked variation in codon usage (Grantham *et al*. 1981). A modulation of the coding strategy according to expression was proposed, such that codons found in abundant mRNAs were under selection for optimal codon-anticodon pairing (Grantham *et al*. 1981). A later codon usage analysis of 83 *E. coli* genes found that variation in codon usage was dependent on translation levels, and that codon usage of abundant protein genes could be distinguished from that of other *E. coli* genes (Gouy and Gautier 1982). Genes with a high protein copy number used a higher frequency of intermediate energy codons and codons that required fewer tRNA discriminations per elongation cycle (Gouy and Gautier 1982).

The distribution of codon bias in *E. coli* was initially reported as bimodal (Blake and Hinds 1984), but it is now accepted that the distribution is unimodal, which presumably reflects a continuum of expression levels (Holm 1986; Ikemura 1985; Sharp and Li 1987a). The distribution of codon bias in *S. cerevisiae* as calculated by the codon bias index and cluster analysis of codon usage, was also described as bimodal (Sharp and Li 1987a; Sharp *et al*. 1986). The original clear distinction between highly and lowly expressed genes was not as apparent in a later analysis but variation in the usage of optimal codons remained the main source of heterogeneity among *S. cerevisiae* genes (Sharp and Cowe 1991).

Codon usage differs between species not only in the selection of codons but in the degree of bias. *B. subtilis* has less biased codon usage than *E. coli*, presumably reflecting a weaker selection, perhaps due to its different environment affecting its effective population size (Moszer *et al*. 1995; Ogasawara 1985; Shields and Sharp 1987). On the other hand codon bias in *S. cerevisiae* is much stronger than in *E. coli* (Sharp *et al*. 1993). Difference in codon bias of homologous genes does not necessarily imply a difference in the expression levels, but rather, it suggests that the effectiveness of the selective pressures on codon usage are not the same.

1.2.1 Co-adaptation of tRNA Abundance and Codon Bias

Ikemura (1981a, 1981b, 1982, 1985) demonstrated that in *E. coli, Salmonella typhimurium,* and *Saccharomyces cerevisiae* codon bias was correlated with the abundance of the cognate tRNA. A strong positive correlation was also found between the copy number of proteins and the frequency of codons whose cognate tRNA was most abundant (i.e. optimal codons) (Ikemura 1981a; Ikemura 1982). This correlation was strongest in the most highly expressed genes, which almost exclusively used “optimal” codons (Ikemura 1981a; Ikemura 1981b; Ikemura 1982; Ikemura 1985), but expression levels and codon choice for *E. coli* plasmid or transposon genes were not found to be significantly correlated (Ikemura 1985). Codon choice at two-fold sites was found to agree broadly with the optimal energy, codon-anticodon interaction theory of Grosjean and Fiers (1982). Ikemura (1982) suggested that bias in codon usage might both regulate gene expression and act as an optimal strategy for gene expression.

1.2.2 Regulation of tRNA Abundance

The co-adaptation of codon usage and tRNA abundance presumably reflects some average growth condition (Berg and Martelius 1995). The total tRNA composition of *E. coli* increases by 50% as growth rate increases to a maximum (Emilsson and Kurland 1990a; Emilsson *et al*. 1993; Kurland 1993), with some tRNA genes being preferentially expressed at high growth rates in *E. coli* (Emilsson, Naslund and Kurland 1993). There are at least two independent regulatory mechanisms for tRNA genes. Some tRNAs are produced at a constant rate relative to cell mass, while others are coupled to the abundance of ribosomes. The tRNAs located in the rRNA operons are used preferentially as major codon species (Komine *et al*. 1990). The rate of the synthesis of these major tRNAs is related to rRNA synthesis, which is in turn related to the growth rate (Jinks-Robertson and Nomura 1987).

Minor codons are not associated with the rRNA operons, although at least one minor tRNA in *E. coli* increases in relative abundance during high growth rates (Kurland 1991). This suggested that it was codon frequency and not the abundance of the cognate tRNAs that determined the response to changes in growth rate (Kurland 1991). Apart from altering tRNA gene frequencies, the nature of genetic variation governing tRNAs is unknown. As a general rule, the major tRNAs are represented as multiple copies in the genome whereas minor tRNAs are represented as a single copy (Komine *et al*. 1990). The over-expression of gene products in *E. coli* produced no specific increases in the relative rates of synthesis of tRNA isoacceptors, but rather a cumulative breakdown of rRNAs and an accumulation of two heat shock proteins, suggesting that the concentrations of many tRNAs are not directly regulated (Dong *et al*. 1995; Nilsson and Emilsson 1994).

1.2.3 Selection for Optimal Codons

Although the correlation between codon frequency and abundant cognate tRNAs was a compelling argument for natural selection choosing between synonymous codons, it could only partially explain the observed bias in codon usage. Ikemura (1981b) described an optimal codon as one that satisfied certain rules of codon choice; the predominant rule is that they are translated by the most abundant cognate tRNA. The rules for the choice of optimal codons were amended and expanded by Ikemura and other investigators as more sequences became available (Bennetzen and Hall 1982; Grosjean and Fiers 1982; Ikemura 1985; Ikemura and Ozeki 1982; Nichols *et al*. 1980).

1.2.4 Translation efficiency

Optimal codons are presumably under selection for some form of translational efficiency and although early *in vitro* measurements of translation rates could find no difference in the rate of translation of optimal and non-optimal codons (Andersson *et al*. 1984), more sophisticated experiments have detected differences in these rates (Sorensen *et al*. 1989). Codons that are recognised by the major tRNAs are translated 3–6 fold faster than their synonyms (Sorensen, Kurland and Pedersen 1989). The rate of initial codon recognition can vary up to 25 fold with optimal codons being recognised most rapidly (Curran and Yarus 1988). It does not necessarily follow that genes that contain a relatively higher proportion of optimal codons, and are presumably translated faster *in vitro*, have higher yields. Perhaps codon usage reflects selection for very fast and short lifetime responses to rapid environment changes (Bagnoli and Lio 1995).

Many of the most highly expressed protein molecules are involved in cell growth and cell division. Rather than optimising the expression of individual genes, codon preferences or "major codon bias" may be part of an overall growth maximisation strategy (Emilsson and Kurland 1990a; Emilsson and Kurland 1990b; Kurland 1991). Kurland (1991) suggested that selection could act upon the translation machinery to improve efficiency, where efficiency implied protein production normalised to the mass of the translation apparatus, the rate of protein production being, most likely, determined by the rate of translation initiation (Kurland 1991). The consequence of faster translation is that ribosomes spend less time on the mRNA, thus elevating the number of free ribosomes and increasing the number of mRNAs translated per ribosome. This is important since the number of ribosomes is often limiting. It has been estimated that up to one third of the dry weight of a rapidly growing *E. coli* cell is ribosomal RNA and protein, and that approximately 70% of the total energy flux in *E. coli* is used in the cellular process of protein synthesis (Ikemura 1985). It has been argued that if protein translation is optimised for the mass invested in the translational apparatus then the translation rate of individual genes cannot be regulated by codon usage (Ehrenberg and Kurland 1984). This is because the average rate at which a particular mRNA is translated will not influence the number of copies of the corresponding protein, since any mRNA represents only a fraction of the overall mRNA pool (Andersson and Kurland 1990).

The three main parameters that affect translation efficiency are (i) the maximum turnover of ribosomes, (ii) the efficiency of aminoacyl-tRNA matching and (iii) ternary complex concentrations (Kurland 1991). Under this model the most efficient mechanism would presumably be the assignment of a single tRNA for each codon or amino acid and this may be analogous to the use of a reduced subset of codons to code abundant proteins and the adjustment of concentrations of individual tRNAs to this pattern. (Ehrenberg and Kurland 1984). The total mass of initiator factors and aminoacyl tRNA synthetases is negligible relative to the masses of the ribosomes. Analyses of isoacceptor concentrations suggest that at low growth rates in *E. coli* the ternary complexes are well below saturation of the ribosome (Kurland 1991). If the abundance of tRNA isoacceptor species match codon bias, the efficiency of translation is enhanced by minimising the mass of aminoacyl-tRNA-GTP-EF-Tu ternary complexes. The overall tRNA/ribosomal ratio decreases with increased growth rate. Evidently the translation apparatus is both expanded and trimmed as growth rates increase; the faster the bacteria grow the more efficiently the mass invested in the translation system is used (Kurland 1993). This increased efficiency is in part due to the reduction in the amounts of tRNA, EF-Tu, and EF-G per ribosome. Accordingly, major codon bias is an aspect of genomic architecture that is selected by the physiological needs of rapidly growing cells (Kurland 1993).

1.2.5 Accuracy and Fidelity in Translation

Selection for fidelity may also be linked to expression level (Gouy and Gautier 1982). It has been estimated that in *E. coli* the non-optimal Asn codon AAU can be mistranslated eight to ten times more often than its optimal synonym AAC (Parker *et al*. 1983; Percup and Parker 1987). Similarly in some contexts the non-optimal codon UUU (Phe) is frequently misread as a leucine codon (Parker *et al*. 1992).

An analysis of the usage of synonymous codons found strong evidence that in highly expressed *Drosophila* genes codon bias is, at least partially, caused by a selection for translational efficiency (Sharp and Matassi 1994). Akashi (1994) examined the codon usage of 38 homologous genes from *Drosophila melanogaster*, *D. pseudoobscura*, and *D. virilis* and found that in genes with weak codon bias, conserved amino acids had higher codon bias than non-conserved residues. In regions encoding important protein motifs (homeodomains and zinc-finger domains), the frequency of preferred codons was higher than in the remainder of the gene, and is was suggested that selection for translational accuracy caused this bias (Akashi 1995). However, a counter argument to this accuracy hypothesis was that the rates of synonymous and non-synonymous substitutions in the homologous genes were not significantly correlated (Akashi 1994). Another counter argument involves the *adh* and *adhR* genes, which encode similar but divergent gene products. Although their primary amino acid sequences have a similar level of conservation between different lineages, *adh* had a strong codon bias and a low K_s (synonymous mutation rate) while *adhR* had a low codon usage bias and a high K_s (Sharp and Matassi 1994). Akashi (1995) found that selection for translational efficiency could influence the observed codon bias in highly expressed *Drosophila* genes. It may be that both translational efficiency and translational accuracy are important in *Drosophila* (Sharp *et al*. 1995). An investigation of homologous *E. coli* and *S. typhimurium* genes found no significant differences in the bias of codons encoding conserved and non-conserved amino acids (Hartl *et al*. 1994).

1.2.6 Rare Codons and Codon context

While some codons are preferentially used in highly expressed genes, some codons are almost absent. These codons are referred to in the literature as rare, unfavoured, or low usage codons. The clustering of rare or unfavoured codons near the start codon was first identified by Ikemura (1981b) in the highly expressed ribosomal protein genes *rplK*, *rplJ,* and *rpsM*. This was attributed to some functional constraint, perhaps a signal for special regulation (Ikemura 1981a). The rarest *E. coli* codons AGA and AGG occur preferentially in the first 25 codons (Chen and Inouye 1994) and in *E. coli* the codon adaptation index (CAI) and synonymous substitution rate of sequence windows are correlated with distance from the initiation codon (Bulmer 1988; Eyre-Walker and Bulmer 1993). However there is not a similar variation in CAI along *B. subtilis* genes (Sharp *et al*. 1990). The bias of conserved codons is also much higher in first 100 codons of homologous genes from *E. coli* and *S. typhimurium*, than in the remainder of the gene (Hartl, Moriyama and Sawyer 1994).

Codons are sometimes found in specific contexts. *E. coli* utilises codon pairs in a non-random pattern (Gutman and Hatfield 1989). Strong correlations between nucleotides at codon interfaces and between wobble positions of adjacent codons suggest that the degeneracy of the genetic code is exploited to arrange codons in some optimal context (Curran 1995). Codon contexts are quite different in highly and lowly expressed genes (Gouy 1987; Shpaer 1986; Yarus and Folley 1985). Though codon contexts seem to be weaker than mutational biases, they may effect observed codon bias; i.e. a gene with a completely optimal synonym choice may not consist entirely of “optimal” codons.

It has been suggested that these observations may be partially a result of avoidance of mRNA secondary structure or additional rRNA binding sites (Bulmer 1988; Eyre-Walker and Bulmer 1993; Hartl, Moriyama and Sawyer 1994). Secondary structure is avoided around the initiation codon of mRNAs (Ganoza and Louis 1994; Wikstrom *et al*. 1992) where it can effect the initiation of translation (de Smit and van Duin 1994; van de Guchte *et al*. 1991). There is also evidence of additional pairing between mRNA and the ribosome after initiation of translation (Petersen *et al*. 1988; Sprengart *et al*. 1990). Even single base pair substitutions that increase secondary structure in the initiation region can have very strong inhibitory effects (500 fold) on the initiation of translation (de Smit and van Duin 1990b). The presence of stable secondary structures in mRNA were not found to cause any appreciable delay in translation; but mRNA levels were reduced 10 fold, presumably the secondary structures were targeted by mRNA degrading enzymes (Sorensen, Kurland and Pedersen 1989).

An analysis of gapA and ompA genes from 10 genera of enterobacteria found a strong bias in their codon usage and surprisingly that different synonymous codons were preferred at different sites in the same gene (Maynard Smith and Smith 1986). Site specific preferences for unfavoured codons were not confined to the first 100 codons and were often manifest between two codons utilising the same tRNA. It was proposed that this was the result of sequence-specific selection rather than sequence-specific mutation (Maynard Smith and Smith 1986).

1.3 Mutation Biases and Codon Usage

Base composition is the most frequently reported DNA feature and is probably one of the most pervasive influences on codon usage. There is wide variation in the genomic G+C content of prokaryotes, ranging from less than 25% to more than 75% G+C content. The G+C content of synonymous third positions can vary by a factor of 10 between species; this bias is always in the direction of the mutational bias. Base composition is a balance between mutational pressure towards or away from G+C nucleotide pairs (Sueoka 1962). The origin of such compositional constraints (GC/AT pressures) is still a matter of debate. Either these compositional constraints are the results of mutational biases (Sueoka 1988; Wolfe *et al*. 1989), or natural selection plays the major role leading to preferential fixation of non-random dinucleotide and base frequencies (Bernardi 1993b; Bernardi and Bernardi 1986; Nussinov 1984). Almost all organisms are subject to directional mutational pressure, and in the absence of selection it is this pressure that shapes gene codon usage (Nichols *et al*. 1980; Sueoka 1988).

Dinucleotide composition also has an appreciable effect on codon choice and is genome specific in both eukaryotes and prokaryotes. For instance dinucleotide TpA appears to be almost universally avoided (Grantham *et al*. 1985) and in many vertebrates the dinucleotide is CpG is relatively rare (Bird 1984). The frequency of a dinucleotide is usually positively correlated with the frequency of its compliment indicating that these biases are characteristics of double-stranded DNA rather than coding mRNA (Nussinov 1981; Nussinov 1984).

The oligonucleotide frequencies in *E. coli* (Phillips *et al*. 1987b) and *S. cerevisiae* (Arnold *et al*. 1988) were found to be much more complex than predicted by the simple over- and under-representation of oligonucleotides and varied in a phylogenetically related way (Grantham, Gautier and Gouy 1980a; Karlin and Cardon 1994; Phillips *et al*. 1987a). Analysis of large genomic regions in both prokaryotic and human sequences, using Markov chain analysis, found regions in many genomes that were atypical. This may be due to unknown selective pressures, structural features or horizontal gene transfer (Scherer *et al*. 1994).

The non-random characteristics of DNA sequences greatly complicate statistical modelling of large genomic DNA sequences (Scherer, Mcpeek and Speed 1994). These patterns include 3rd codon position periodicity (Lio *et al*. 1994), a universal G-non-G-N codon motif (Trifonov 1987), and long-range power-law correlations (Ossadnik *et al*. 1994). Statistical analysis of eukaryotic DNA sequences using techniques derived from linguistics, found that non-coding sequences have characteristics that are similar to natural language, with smaller entropy and larger redundancy than coding sequences. This has been interpreted as evidence that “non-coding” sequences carry biological information, which is perhaps not surprising (Mantegna *et al*. 1994).

1.3.1 Variation of Codon Usage with Genome Location

Mammalian genomic DNA, originally thought to have quite a narrow range of G+C content (Sueoka 1961), has large regional differences in base composition. These relatively long tracts of DNA (300 kb), which differ in their local G+C content, are termed isochores (Bernardi 1989; Bernardi 1993b; Bernardi *et al*. 1985). The origins of isochores are still shrouded in some mystery (Sharp and Matassi 1994). Isochores have been classified into two light or AT rich classes (L1 and L2) and three heavy G+C rich classes (H1, H2 and H3) (Bernardi 1993b). Gene density is non-uniform; low G+C isochores (L1 and L2) comprise approximately 60% of the human genome, but only one third of genes lie within these regions. A third of genes lie within H3, which only comprises 3–5% of the genome (Bernardi 1993a). There has been a suggestion that some housekeeping genes may be located preferentially in H3 isochores (Bernardi 1993b). Hybridisation of the human H3 isochore with other mammalian and avian genomes has shown that the structure of isochores is conserved remarkably well between species (Caccio *et al*. 1994).

Among 20 mammalian species belonging to nine different eutherian orders, only the myomorph rodents (mouse, rat, hamster, and mole rat), the pangolin, and the fruit bat were found to differ from the ‘general’ pattern (as found in humans). This was principally because they lacked the most G+C rich H3 isochores (Sabeur *et al*. 1993). It is not clear how these patterns have diverged (Bernardi 1993a). Avian species contain isochores and because birds speciated from the mammalian orders before reptiles, this has been interpreted as evidence for at least two independent origins of isochores (Bernardi 1993a).

The pattern of codon usage in angiosperms indicates that they may also contain isochores (Matassi *et al*. 1989). Variation in the G+C content of silent sites is the major source of variation in codon usage (Fennoy and Baileserres 1993). It is difficult to identify whether this GC_{3s} base variation is due to regional effects or translational selection. Codon usage has been reported as being more biased in some highly expressed chloroplast genes, histones and anthocyanin biosynthetic enzymes (Fennoy and Baileserres 1993). The main difference in the codon usage between monocotyledons and dicotyledons is the average GC_{3s} of the genes. Those genes expected to be highly expressed are reported as having a more biased codon usage than genes expected to be moderately or lowly expressed (Murray *et al*. 1989; Tyson and Dhindsa 1995).

In mammals codon usage varies enormously among genes (Mouchiroud and Gautier 1990; Newgard *et al*. 1986). However, this probably only reflects the general phenomenon of G+C variation with location (Ikemura and Wada 1991) as there is scant evidence that it has been shaped by selection for translation efficiency (Sharp *et al*. 1993). There is a correlation between gene density and G+C content, but the location of genes appears to be independent of tissue, time or level of gene expression (Bernardi 1993b). There is also a correlation between the G+C content of the 1^{st} and 3^{rd} codon positions of mammalian genes (Eyre-Walker 1991). Patterns such as a preference for pyrimidine-purine codon boundaries also influence the observed codon bias (Galas and Smith 1984; Smith *et al*. 1985).

The substantial GC_{3s} variation between prokaryotic genes has been used to infer the presence of isochores in prokaryotes (D'onofrio and Bernardi 1992; Sueoka 1992), but the influence of translational selection which is known to influence GC_{3s} strongly was ignored. There is only enough sequence information to ask whether gene location influences codon usage for a small number of prokaryotic species (Sharp and Matassi 1994). Genes that have weak codon bias display GC_{3s} variation that is associated with chromosomal position, with a lower GC_{3s} near the terminus of replication (Deschavanne and Filipski 1995). In *E. coli* chromosomal location influences substitution rates, genes located near the origin of replication have lower substitution rates (Sharp *et al*. 1989). This implies that either mutational biases or natural selection vary systematically with genomic location. The mechanisms by which location can influence gene evolution have received far less attention than the effect of natural selection on synonymous codon usage (Sharp and Matassi 1994). These regions of differing GC_{3s} have been termed chichores (Deschavanne and Filipski 1995).

While the codon usage of *S. cerevisiae* had been extensively quantified (Ikemura 1982; Sharp and Cowe 1991; Sharp *et al*. 1988), the publication of the complete sequence of the *S. cerevisiae* chromosome III (Oliver *et al*. 1992), allowed codon usage variation to be examined as a function of chromosomal location. Chromosome III is approximately 315 kb long, with the right arm slightly longer than the left (Oliver *et al*. 1992). Genes that are G+C rich at silent sites (i.e. with a high GC_{3s}) are located predominantly in two distinct chromosome regions. These approximate to the centre of the two chromosome arms (Sharp and Lloyd 1993), while regions poorer in G+C are found at the centromere and telomeres. This G+C variation is independent of the selection for optimal codons (only half of the *S. cerevisiae* optimal codons end in G or C). Multiple periodic G+C peaks were also reported for the next three chromosomes XI, II and VIII (Dujon *et al*. 1994; Feldmann *et al*. 1994; Johnston *et al*. 1994), with approximately one peak per 100 kb (Sharp *et al*. 1995). A correlation between silent site G+C (GC_{3s}) content and gene density was noted during the analysis of chromosome XI (Dujon *et al*. 1994), this correlation was also found in chromosome III (Sharp and Matassi 1994) and in the subsequent primary publications for chromosomes II, IV, VIII, XIII, and XV (Bowman *et al*. 1997; Dujon *et al*. 1997; Feldmann *et al*. 1994; Jacq *et al*. 1997; Johnston *et al*. 1994). However, a recent analysis of all *S. cerevisiae* chromosomes found that there was no correlation between gene density and GC_{3s} (Bradnam *et al*. 1999). While variation in GC3s is not completely random the observed clusters of ORFs of similar GC3s can be accounted for by very short-range correlations between neighbouring ORFs. Bradnam *et al.* (1999) also reported that high G+C ORFs are located preferentially on shorter chromosomes and that in many ways chromosome III was atypical of the other chromosomes.

In *Borrelia burgdorferi* it is a genes' orientation relative to direction of DNA replication, not its location on chromosome, which determines is codon usage pattern (McInerney 1998). An analysis of the genomes of spirochaetes *Borrelia burgdorferi* (McInerney 1998) and *Treponema pallidium* (Lafay *et al*. 1999) found that there was no evidence for translation selection operating on the codon usage of highly expressed genes. Codon and amino acid usage composition patterns differ significantly between genes encoded on the leading and lagging strands.

1.3.2 Time of Replication

The mechanisms that cause G+C mutation patterns to vary have been the subject of considerable debate. At particular issue is whether these isochores are in some way adaptive or are the passive result of mutational processes. Kadi *et al*. (1993) explain the presence of isochores in warm-blooded animals as resulting from natural selection, but the mechanism by which this occurs is elusive. Other investigators prefer the hypothesis that variation in G+C content arises because the isochores are replicated at different points of the cell replication cycle. If the G+C content of the nucleotide pools varied, they would presumably affect mutation bias (Wolfe, Sharp and Li 1989). This hypothesis has been supported by recent models of the origin of isochores (Gu and Li 1994). There is probably more detailed information about the replication of the first 200kb of yeast chromosome III than any other eukaryotic chromosome (Sharp and Matassi 1994). Despite this, no obvious relationship has been found between replication timing and G+C content (Dujon *et al*. 1994).

1.3.3 DNA Repair

It has been suggested that the efficiency of DNA mismatch repair mechanisms might vary with chromosomal location (Filipski 1987; Hanawalt 1991) but this would presumably leave a signal, in the form of a strong correlation between G+C content and substitution rates. This signal is not seen, suggesting that codon usage can only be explained in terms of a variation in DNA mismatch repair under a restricted set of circumstances (Eyre-Walker 1994a). Genes which are transcribed more often (i.e. highly expressed genes) may have lower mutation rates because they are subject to a more rigorous DNA repair response (Berg and Martelius 1995). The coupling of the repair of pyrimidine dimers with transcription has been identified in *E. coli* (Selby and Sancar 1993). The RNA polymerases pause at the pyrimidine dimers and this signals the repair machinery (Friedberg *et al*. 1994). It has been suggested that the efficiency of the very short repair mechanism changes with codon bias/gene expression (Gutierrez *et al*. 1994) but this correlation seems to be an artefact of codon bias as codon CTA is rare in *E. coli* and the codon TAG is absent which may go some way to explain the rarity of CTAG (Eyre-Walker 1995b).

1.4 Codon Usage

1.4.1 Optimal Codons

When Ikemura (1985) defined the optimal codons in *E. coli*, *S. typhimurium*, and *S. cerevisiae*, his definition was dependent on knowledge of the abundance and characteristics of their tRNA molecules. The number of species where the abundance and structures of tRNAs are known is limited relative to the number of organisms from which sequence data has been obtained. Indeed, what knowledge there is of tRNA abundance is potentially biased, because measurements are made under laboratory growth conditions. It is therefore desirable to define an optimal codon in terms of a more readily estimated characteristic. The most commonly used characteristic is the pattern of codon usage itself, the definition used in this thesis is “an optimal codon is any codon whose frequency of usage is significantly higher in putatively highly expressed genes” (Lloyd and Sharp 1991; Lloyd and Sharp 1993; Sharp and Cowe 1991; Sharp *et al*. 1988; Shields and Sharp 1987; Stenico *et al.* 1994). Significance is estimated using a two-way chi-squared contingency test, with a cut-off at $p<0.01$. The most frequent codon for an amino acid is not necessarily an optimal codon, which is subtly different from the original definition of an optimal codon used by Ikemura (1981b), who defined optimal codons as those codons occurring most often in biased genes.

1.4.2 Mutation Selection Drift

Codon usage variation is represented by two major paradigms. Either mutational bias and selection determine codon usage, or it is determined by mutational bias alone. Although natural selection for efficient translation is a major influence on codon usage in many species, it is not always apparent in what form the selection is taking place and it does not explain all of the observed codon usage variation. Some genes have codon usage that is determined mainly by mutation and drift while others display codon usage that arises from a balance between mutational biases and selective pressures (Berg and Martelius 1995; Bulmer 1988; Sharp and Li 1986). Observed codon bias is an equilibrium between selection that favours the fixation of advantageous codons and genetic drift that enhances the probability of the fixation of disadvantageous codons (Akashi 1995; Bulmer 1988).

While the development of a unified theory for codon usage has so far proved elusive, the mutation-selection-drift (MSD) theory has been described as a reasonable working hypothesis (Bulmer 1991) and is the most widely accepted (Akashi 1995; Hartl, Moriyama and Sawyer 1994; Kurland 1993; Sharp *et al*. 1993).

The maintenance of codon preference for nearly neutral synonymous positions requires a slow but constant rate of adaptive fixation (Akashi 1995). If selection acts independently on each codon then selective differences between synonyms are probably very small, so codon selection will only be effective in species with very large population sizes (Bulmer 1991; Li 1987). Selection is likely to be stronger in highly expressed genes because these codons are translated more often (Bulmer 1988). Bulmer (1991) suggested that selection coefficients for optimal codons, based on protein expression levels, would be in the order of 10^{-4}, but this value appears to be rather high as it implies that the *E. coli* N_e would be of the order of 10^4, which is much lower than estimates of N_e in *E. coli* (Hartl, Moriyama and Sawyer 1994). Other estimates for codon selection coefficients in *E. coli* have been of the order of 10^{-8} (Akashi 1995; Hartl, Moriyama and Sawyer 1994). The product of the effective population and selection coefficient N_es for disfavoured synonymous codons in the highly expressed *gnd* and *putD* has been estimated as approximately -1.3 (Hartl, Moriyama and Sawyer 1994). In the *E. coli gnd* (6-phosphogluconate dehydrogenase) the selection against detrimental codons has been estimated as one third of the selection coefficient against detrimental amino acid replacements (Hartl, Moriyama and Sawyer 1994).

1.4.3 Mathematical Models

Within the framework of the neutral mutation-random drift theory, Kimura (1981) proposed that random drift around some optimum value (under stabilising selection) could explain the observed non-random or unequal usage of synonymous codons. Li (1987) felt that directional selection, rather than stabilising selection, was the most appropriate assumption for a codon usage model. Constant selection models require a very restricted range of selection intensity to explain the observed codon bias (Eyre-Walker 1994b). Akashi (1995) in turn has suggested that the synergistic model might be necessary to explain the available data where species with effective population sizes that differ by several orders of magnitude appear to have similar degrees of codon bias. Kimura (1981) felt that the most plausible explanation for preferential codon usage was that it represented an optimum state, where the choice of synonymous codons matched the cell’s cognate tRNAs concentrations. This would reduce substitution rates at silent sites to maintain a given optimum equilibrium bias (Kimura 1983).

Li (1987) described intermediate codon bias as a balance between genetic drift and selection, and described the relative frequencies of synonyms as a function of mutational bias, selection coefficient (s), and effective population size (N_e). To select between synonyms, the selective advantage must be greater than the inverse of the effective population size. Synonymous sites would be fixed when the absolute rate of mutation was low and the effective population size was small, such that population polymorphism would be negligible. If N_es fell much below unity, drift would overwhelm codon usage. If it rose above three or four (depending on the mutational bias), all codons would be fixed for the preferred codon. This assumes independent segregation of codons; linkage would substantially increase the accumulation of slightly deleterious codons.

Bulmer (1987) combined aspects of previous translation models and investigated how bias may develop in organisms with a large enough population size, due to selection for translational efficiency. When selection was greater than the mutation rate, codon usage would co-adapt with the translational machinery such that the number of tests of non-cognate tRNAs would be minimised. This model has been described as unrealistic (Shields 1989). It assumed that the population size would be large enough to suppress the effect of stochastic fluctuations caused by random genetic drift, which would randomise codon frequencies. Under this model, the continued presence of disadvantageous codons was due to the continual occurrence of mutations in the population. These would be unlikely to become fixed, as selection would eliminate them from the population before their frequency became too great. A disadvantage of this model was that it predicted very high sequence polymorphism in lowly biased genes and, as long as codon preference was maintained, the absence of sequence divergence at silent sites over evolutionary time. Obviously DNA sequences have diverged and this must be caused by the fixation of mildly deleterious alleles, implying that the effective population must be finite. This model was later enhanced to take into account population size and selective differences between codons (Bulmer 1991). However when this newer model was tested, Bulmer (1991) found that it "grossly overestimated codon bias" in highly expressed genes (ribosomal protein and AA-tRNA synthetase genes).

Shields (1989) proposed a model for codon usage where selection, mutational bias, and effective population size shaped codon preference. Codon usage was dependent on both the magnitude and variability of selection pressures. The frequency of an optimal codon in highly expressed genes would be largely insensitive to changes in mutation bias, unless the bias exceeded a critical value. This could then result in a switch of optimal codon (Shields 1989). When selection was stronger, a stronger mutation bias against a codon would be required to alter it. This contradicted a previous prediction that selection pressures and mutational biases acted additively in highly expressed genes to influence codon usage (Osawa *et al*. 1988).

As the magnitude of species' effective population size varies considerably (Nei and Graur 1984) it seems probable that it has also fluctuated during evolution. Under Shield's (1989) model a decrease in population size could result in the selection for synonyms no longer being effective, such that codon usage would be entirely determined by mutational biases. A codon that was advantageous but not favoured by mutational bias could be replaced in abundance by a synonym that was more favoured by mutational bias. If the population size increased, the tRNA population would co-adapt with the more abundant codon and thus the optimal codon could switch. Changes in mutation patterns may be the major cause of switches in codon preferences (Shields 1990). This model of codon usage assumes that the tRNA population adapts to the more frequent codons, in such a way that they are translated more efficiently. Analysis of the codon usage of Enterobacteria indicated that the observed data was largely consistent with this model (Shields 1990). It also explained why the highly expressed genes in *S. marcescens* and *E. coli* have similar GC_{3s}, in contrast to the lowly expressed genes which have a much higher GC_{3s} in *S. marcescens* than in *E. coli* (Sharp 1990). Since the divergence of these species there has been little change in the optimal codons despite differences in mutation bias (Shields 1990). In *Proteus vulgaris* the optimal codons have diverged, reflecting that mutational bias has been strong enough to precipitate switches in codon preference.

The models of Bulmer (1991) and Li (1987) provided useful limiting cases but a problem with Bulmer's (1991) model was that it overestimated the predicted frequency of rare codons in highly expressed genes. If any of the selection coefficients were less than the inverse of the effective population size, codon usage would be randomised by genetic drift. If selection coefficients were less than mutation rate, recurrent mutation would prevent selective codon usage evolving. The Shields (1989) model described how a change in mutation patterns or selective pressures or population sizes could change codon usage. While models of codon usage are useful tools for exploring the mechanisms by which the tRNAs and codon usage may have adapted to the "problem" of optimising translational efficient it is important to realise that these mechanisms are much more complex than current models can allow for.

1.4.4 Codon Usage Patterns

1.4.4.1 Prokaryotes and Unicellular Eukaryotes

Though understanding of codon usage is more advanced for the prokaryotes than for the eukaryotes, much of this knowledge is based on the relatively few species that have been subjected to a concerted molecular genetic analysis. Our understanding of codon usage among the Gram-negative proteobacteria is much more advanced than in any other group of species. The codon usage of the model organism *E. coli* has been extensively investigated (Gouy and Gautier 1982; Grantham, Gautier and Gouy 1980a; Grosjean and Fiers 1982; Ikemura 1981a; Ikemura 1981b). In the Gram-positive bacteria (with the exception of *Bacillus subtilis*), an understanding of codon usage patterns has been severely limited by the lack of sequence information. The importance of an adequate sample size in the analysis of codon usage cannot be over emphasised. An analysis of *B. subtilis* codon usage based on only 21 genes reported that all codons were used more or less equally (Ogasawara 1985) but later analyses with a greater number of genes reported translational selection among synonyms (Sharp *et al*. 1990; Shields and Sharp 1987).

In many prokaryotes with extreme G+C mutational bias, GC_{3s} is often so biased that functional open reading frames are easily recognisable (Bibb *et al*. 1984). In the A+T rich Gram-positive *M. capricolum*, ribosomal proteins have a very high frequency of codons ending in A or T (Muto *et al*. 1984; Muto *et al*. 1985). Conversely the G+C rich *Thermus thermophilus* has a high frequency of codons ending in G or C (Kagawa *et al*. 1984; Kushiro *et al*. 1987). In some prokaryotes, particularly those with G+C rich or G+C poor genomes (e.g. *Mycoplasma capricolum*, *Micrococcus luteus*, and *Streptomyces* species), if natural selection is choosing between synonymous codons it is much weaker than the influence of mutational bias and is swamped by the latter (Sharp *et al*. 1993). *Mycobacterium tuberculosis* and *Corynebacterium glutamicum* are both G+C rich Gram-positive bacteria, and although neither is extremely biased in base composition putative translationally optimal codons have been identified in both species (Andersson and Sharp 1996; Malumbres *et al*. 1993). An exception to the generalisation that genomes with extreme genomic G+C biases do not display codon preference is the codon usage of *Dictyostelium discoideum* (overall G+C content of 22%) (Sharp and Devine 1989). Codon usage in *D. discoideum* reflects its A+T richness but a subset of codons (mainly C ending) appears to be translationally optimal. Some of these codons (UUC, UAC, AUC, AAC, GAC, GGU) are also optimal in *E. coli*, *B. subtilis*, *S. cerevisiae*, *S. pombe* and *D. melanogaster* and these codons have been described as universally optimal (Sharp and Devine 1989).

While in almost all lineages the genetic code has remained constant, codon usage and the choice of optimal codons have diverged. A detailed and accurate analysis of codon usage is an essential prerequisite to our understanding of how and why divergent patterns of codon choice evolved. There is no obvious reason why the subset of optimal codons should differ between species. Codon usage (i.e. the choice of optimal and non-optimal codons) is broadly similar in closely related species but diverges with increasing phylogenetic distance. The codon usage of *S. typhimurium* is the same as that of *E. coli* (Sharp 1991), which may simply be because an insufficient number of substitutions have occurred for a difference to be detected. The similarity in the codon usage biases of homologous genes has been used to suggested that the selection pressure on synonymous codons has been similar since the species diverged 10^8 years ago (Ochman and Wilson 1987a). The codon usage and tRNA population of the more distantly related species *S. marcescens* (Ochman and Wilson 1987a) have also remained similar to those of *E. coli* (Ikemura 1985; Sharp 1990). The total codon usage and the choice of optimal codons of the Gram-positive species *Bacillus subtilis* are distinct from those of *E. coli*, the overall codon usage and choice of optimal codons has altered to the extent that AT-rich codons predominate in *B. subtilis*, reflecting its lower genomic G+C content. However, many codons remain optimal in both species (Moszer, Glaser and Danchin 1995; Ogasawara 1985; Shields and Sharp 1987). Within the phylogenetically diverse genus *Lactobacillus*, codon usage bias is correlated with expression, but varies between species (Pouwels and Leunissen 1994).

Between *E. coli* and *S. cerevisiae* the most abundant tRNAs differ for approximately half of the amino acids, and there is a correlated change in the choice of optimal codons (Ikemura 1985).

The choice of optimal codons is the same for the distantly related *Kluyveromyces lactis* and *S. cerevisiae* despite the saturation of their silent sites, which presumably arises from a similarity in the mutational biases and underlying tRNA pools of *K. lactis* and *S. cerevisiae* (Lloyd and Sharp 1993). Codon bias for some genes differs between these two species, but in a manner correlated with differences in expression level (Freirepicos *et al*. 1994). The codon usage of *S. cerevisiae,* the distantly related ascomycete fungi *Aspergillus nidulans*, and *Schizosaccharomyces pombe* have however diverged (Lloyd and Sharp 1991; Sharp and Wright 1988).

Although codon usage changes over evolutionary time, the similarity of parameters that constrain codon usage can cause convergence in distantly related organisms (Sharp and Cowe 1991). When examining codon usage it is important to distinguish between interspecific and intraspecific variation. In addition, it is necessary to consider whether the variation is caused by a mutation bias or a selection for a translationally efficient codon dialect. For example, the G+C rich *Serratia marcescens* (59% genomic G+C content) has a high variation in GC_{3s} (G+C content at synonymous third positions) values. Although this has been attributed to a variation in genome mutation bias (Nomura *et al*. 1987), it is more readily explained as an equilibrium between mutation and selection (Sharp 1990).

1.4.4.2 Multicellular Eukaryotes

Despite there being striking differences in codon usage and codon bias of mammalian genes, there is no codon usage preference in human genes *per se* (Bernardi 1993a; Ohno 1988; Sharp and Matassi 1994). Differences in codon choice can be attributed to variation in the GC_{3s} of mammalian genes. The GC_{3s} of mammalian genes is strongly correlated with the G+C content of introns, 5’ and 3’ sequences (Andersson and Kurland 1990; Aota and Ikemura 1986; Bernardi 1993a; Ikemura 1985; Sharp *et al*. 1993), with neighbouring genes have similar GC_{3s} values (Ikemura and Wada 1991). It was thought that there was a fundamental dichotomy between the codon usage of unicellular and multicellular organisms, with the codon usage of the unicellular eukaryotes (e.g. *Saccharomyces cerevisiae, Schizosaccharomyces pombe*) and prokaryotes being determined by mutation-selection-drift, and that of multicellular organisms by mutational bias and drift (Ikemura 1985). However, *Drosophila melanogaster* and

Caenorhabditis elegans display a bias in their choice of codons which appears to be caused by selection for translation efficiency (Shields *et al*. 1988; Stenico, Lloyd and Sharp 1994). The absence of selection between synonyms in mammals is not simply a result of the subdivision of multicellular organism into different cell types.

The codon usage of *Drosophila melanogaster* is more similar to the *E. coli*/yeast paradigm, than to that of mammals (Moriyama and Hartl 1993; Shields *et al*. 1988). It seems that the subtle differences between one synonym and another have a discernible effect on the chance of a fruit fly surviving and reproducing (Sharp and Matassi 1994). Codon bias appears to be maintained among close relatives of *D. melanogaster*, such that genes with high codon bias exhibit less divergence at silent sites (Moriyama and Gojobori 1992; Sharp and Li 1989). Isochores, which dominate mammalian codon usage, were not found in *D. melanogaster* (Bernardi *et al*. 1985).

Histone genes are an exception to the trend of highly expressed Drosophila genes having biased codon usage. They are highly expressed and highly conserved proteins, but have low codon usage bias and a relatively high rate of synonymous substitution (Fitch and Strausbaugh 1993). Natural selection has great difficulty in distinguishing between the best variants at multiple sites if these sites are tightly linked (Kliman and Hey 1994). The rate of recombination is much reduced in regions near the telomeres and around the centromeres. Genes located in these regions, which include the histone genes, have lower codon bias (Kliman and Hey 1993).

Nei and Graur (1984) have estimated the effective population size for *Drosophila* to be between 10^6 and 10^7, while similarly derived N_e values for mammals are of the order 10^4. If the selective coefficients for codons in these eukaryotes are in the order of 10^{-5} and 10^{-6}, this would account for the presence of selective bias in *D. melanogaster*. Since *D. melanogaster* and *D. simulans* diverged from their common ancestor there has been an apparent relaxation in selection at silent sites and in codon bias; *D. melanogaster* has an estimated absolute N_es of approximately one (Akashi 1995). The N_es of highly expressed *D. simulans* genes have been estimated by Akashi (1995) to be approximately 2.2; however the N_e for these species has been estimated to vary 20-fold (Hartl, Moriyama and Sawyer 1994; Nei and Graur 1984). This apparent relaxation is further supported by estimates of DNA heterozygosities (Aquadro 1992) although differences in selection intensity at different growth rates may also have an influence.

Synonymous codon usage varies considerably among *C. elegans* genes, with a single major trend in the variation. The frequency of a subset of codons appears to be correlated with the level of gene expression (Stenico, Lloyd and Sharp 1994). There has also been a great deal of interest in the codon usage of parasitic helminths and nematodes (*Brugia*, *Echinococcus*, *Onchocerca,* and *Schistosoma*). Codon bias has been reported but there is no evidence for translational selection (Ellis *et al*. 1993; Ellis *et al*. 1995; Ellis *et al*. 1994; Kalinna and McManus 1994). Analysis of the codon usage of *Schistosoma mansoni* found that bias was dependent on the overall base composition of the genes analysed (Ellis and Morrison 1995; Milhon and Tracy 1995; Musto *et al*. 1994).

The codon usage of chloroplast and mitochondrial genomes differ from the codon usage of their host cells in both their rate and patterns of evolution (Bonitz *et al*. 1980; Pfitzinger *et al*. 1987). The codon usage of *psbA,* the most highly expressed gene in the *M. polymorpha* chloroplast, is markedly different from other chloroplast genes. This has been attributed to selection for optimal translation (Morton 1994).

1.4.5 Initiation and Termination Codon Usage

There has been a great deal of interest in the evolution of codon usage around initiation and termination codons, the base composition, sense codon usage, and frequency of amino acids exhibit significant deviations from a random distribution, this is accentuated in highly expressed genes (Alffsteinberger and Epstein 1994; Brown *et al*. 1993; Brown *et al*. 1994; Sharp and Bulmer 1988).

In *E. coli,* the 60-80 nucleotides that bracket the gene initiation codon generally promote translation. This extends beyond the mere presence of a Shine-Dalgarno element followed by a suitable start codon. Some general mechanism must protect these against sequestration by long-range base pairing. A reasonable guess is that the sequence around start codons is constrained to minimise the local structure of the ribosomal binding site to keep translational start sites available to ribosomes (de Smit and van Duin 1990a; Jacques and Dreyfus 1990).

Base composition at silent sites is skewed at the start of genes, the frequency of A is higher and G lower in all three codon positions. Some of the codon bias near the initiation codon can be explained as amino acid selection, the excision of the N-terminal methionine is dependent on the length of the following amino acid’s side chain (Hirel *et al*. 1989). The N-terminal amino acid can have a large effect on the half-life of a protein (Tobais *et al*. 1991).

The three standard termination codons have different properties; a very important one is the propensity to which a termination signal can be read through. The termination codon UAA is the least leaky (Tate 1984), while UGA is most likely to promote translational frame shifts (Weiss *et al*. 1987). The choice of termination codon correlates with gene expression level (Sharp *et al*. 1992). In highly expressed genes there is a strong bias for UAA (which is recognised by two release factors RF-1 and RF-2) (Sharp and Bulmer 1988). The concentrations of RF-1 and RF-2 vary with growth rate in *E. coli*; RF-1 increases from 1,200 to 4,900 copies and RF-2 from 5,900 to 24,900 copies as growth rate increases. Due to the net increase in the cellular mass involved in translation in the cell, this equates to a net 1.5 fold increase in the overall concentration of these release factors (Adamski *et al*. 1994).

Suppressible mutations have shown that termination efficiency is strongly dependent on the 3’ context, so much so that the stop signal has been described as a four base signal (Brown *et al*. 1994). The efficiency of the 12 possible 'four base stop signals' (UAAN, UGAN and UAGN) vary significantly depending on both the stop codon and the fourth base, ranging from 80% (UAAU) to 7% (UGAC) (Poole *et al*. 1995). The rate of release factor selection varied 30-fold at UGAN stop signals, and 10-fold for both the UAAN and UAGN series. This correlates with the frequency that these signals are found in nature. It also provides a rationale for the presence of the strong UAAU signal in many highly expressed genes and the presence of the weaker UGAC signal at several recoding sites (Poole, Brown and Tate 1995). Preferred stop codon contexts are also found in human genes (Martin 1994). These contexts appear similar to those found in *E. coli* (Arkov *et al*. 1995). However it is not clear if the identity of the 3' base is determined by genome wide changes in G+C composition, or selection to maintain a particular tetranucleotide stop signal (Martin 1994).

1.4.6 Horizontal Gene Transfer

It is not yet clear to what extent inter-species recombination occurs among prokaryotes. Gene transfer is often associated with transposon-like elements or insertion sequences (Groisman *et al*. 1992; Groisman *et al*. 1993; Simon *et al*. 1980). Before horizontally transferred sequences can be established, they must overcome transfer barriers that prevent the delivery of genetic information from a donor cell and establishment barriers that block inheritance of newly acquired genes (Matic *et al*. 1995). Genes acquired by horizontal transfer often have atypical G+C content, codon bias and repetitive elements (Medigue *et al*. 1991), and only approach the characteristic codon usage and G+C content of their host after millions of years (Groisman *et al*. 1993).

Medigue *et al.* (1991) applied correspondence analysis and cluster analysis to the investigation of the codon usage of 780 *E. coli* genes, and described three classes of genes. Class III genes having codon usage that does not reflect the average distribution of specific tRNAs, so they have low CAI values. Oligonucleotide analysis indicated that in class III many of the rare oligonucleotides of classes II and I are evenly distributed (Medigue *et al*. 1991). It was concluded that class III genes are mostly comprised of genes that are exchanged horizontally and that they represented a significant fraction of the *E coli* chromosome (Medigue *et al.* 1991). The classes II and I are similar to grouping identified by Gouy and Gautier (1982). The distribution of codons was quite unbiased in class III, for example the rare codon AUA is used for 26% of Ile residues and no codon was used less than 7%. Analysis of genes known to be horizontally transferred such as lambda, plasmid and transposon genes indicated that they clustered with class III genes (Medigue *et al*. 1991).

Perhaps one fifth of *E. coli* genes undergo continuous exchange with other microbial genomes (Borodovsky *et al*. 1995). The majority of genes that have been suggested as candidates for horizontal transfer in *E. coli* are genes whose acquisition presents an immediate adaptive advantage; e.g genes encoding cell surface proteins and antibiotic resistance genes (Matic *et al.*

1994; Matic, Rayssiguier and Radman 1995; Smith *et al*. 1990; Verma and Reeves 1989). Examples include the *lac* operon (Buvinger *et al*. 1984) and the *umuCD* operon, required for mutagenic DNA repair (Sedgwick *et al*. 1988). While the *umuCD* operon is present in both *E. coli* and *S. typhimurium* it is highly diverged between the two species (Sharp 1991).

Other examples of horizontally transferred genes include the O antigen and phosphatase gene of *S. typhimurium* (Groisman, Saier and Ouchman 1992; Reeves 1993), and the *catIJF* operon of *Acinetobacter calcoaceticus* (Shanley *et al*. 1994). Genes involved in antibiotic resistance have been widely horizontally transferred, though this is generally under very intensive selective pressure (Martin *et al*. 1992; Spratt *et al*. 1992).

1.4.6.1 Overlap between E. coli Genes

In *E. coli,* genes with a CAI below 0.45, (i.e. lowly biased genes) are much more likely to have the preceding gene overlap their start codon, most commonly by one or four base pairs. Genes with a CAI greater than 0.45 overlap with the preceding gene infrequently and the preceding gene infrequently terminates within 10 base pairs. Whether this is due to selection in lowly biased or highly biased genes is unclear (Eyre-Walker 1995a).

1.4.7 Codon Usage and Phylogeny

As codon usage divergence is correlated with evolutionary distance (Grantham *et al*. 1981; Long and Gillespie 1991; Maruyama *et al*. 1986), it has been suggested that codon usage (Goldman and Yang 1994; Nesti *et al*. 1995; Pouwels and Leunissen 1994) or amino acid usage (Schmidt 1995) can help unravel the evolutionary relationships between species. Although phylogenies based on codon usage may appear to have practical application, phylogenies are best investigated by comparative analysis of homologous sequences (Sharp 1986).

As rather few genes have been found in many species, it is still not possible to characterise the inter-species variation of codon usage in detail. Codon usage can converge in an evolutionary distant species due to similar mutational bias. A phylogeny of seven species from the phylum

Apicomplexa, based on codon usage divergence; has been used to support the hypothesis that codon usage can be used to estimate phylogenies (Morrison *et al*. 1994). Nesti and co-workers (1995) also presented a phylogeny based on codon usage divergence, however their paper might equally be used as evidence for the drawbacks of such a technique. Although parts of their topology were valid, some were erroneous because the codon usage of distantly related species had converged. For example, the low G+C prokaryotes, *S. aureus,* and *B. subtilis* clustered with the low G+C eukaryotes *Plasmodium falciparum* and *Dictyostelium discoideum,* rather than with the other prokaryotes.

1.4.8 Amino Acid Composition

Though only working with eight *E. coli* genes, Ikemura (1981b) noted a strong positive correlation between amino acid composition and codon bias. This has been shown, after hydrophobicity, to be the second strongest trend in the amino acid composition of *E. coli* (Lobry and Gautier 1994). Surprisingly this is a more significant trend than aromaticity, amino acid volume, or charge (Lobry and Gautier 1994). Many highly expressed proteins are quite basic, presumably because many of them are ribosomal proteins that must interact with DNA. As growth rate increases the basic amino acids Arg and Lys, increase in abundance by 20% and 8% respectively, relative to the total amino acid concentration. The aromatic amino acids Phe and Tyr decrease by 16% and 23 % respectively (Kurland 1991), this is possibly a growth optimisation strategy (Andersson and Kurland 1990).

1.4.9 Complimentary ORFs

The presence of “shadow codons” (Grantham *et al*. 1985) or complimentary codons has been found in both human and *E. coli* genomes (Alffsteinberger 1984). Complementary ORFs in coding sequences are not uncommon, but are probably artefacts of codon usage due to the relative scarcity in real genes of the codons UUA, UCA, and CUA which are complementary to stop codons and due to an excess of RNY codons (Sharp 1985). Randomised sequences have a similar frequency of complimentary codons and higher order oligonucleotides (Forsdyke 1995a; Forsdyke 1995b). It has also been suggested that complimentarily ORFs may be due to the wide distribution of inverted repeats in natural DNA sequences (Merino *et al*. 1994).

1.5 Molecular evolution

The evidence that natural selection could influence silent changes (Grantham *et al*. 1981; Grantham *et al*. 1980b; Ikemura 1985; Kimura 1983), suggested that in some (perhaps many) genes in some (perhaps many) species, silent sites were not neutral (Sharp *et al*. 1993). The corollary of this, is that some synonymous substitutions are effectively neutral and probably accumulate at frequencies approaching the mutation rate (Ikemura 1981a; Ikemura 1981b; Ikemura 1985; Ochman and Wilson 1987b). Analysis of codon usage can infer both the nature and strengths of some of the selective forces to which the organism has been exposed (Sharp and Cowe 1991). They can reveal the rates and patterns of silent site evolution and allow the investigation of how natural selection selects between mutations that (presumably) cause very small differences in fitness (Akashi 1995). Weak selection allows non-adaptive processes to be evident and with a large number of sites under broadly similar constraints, it is possible to perform the quantitative analyse of data. With the identification of advantageous codons it is possible to predict the relative advantage and disadvantage of alternative sequences and perhaps the prediction of a completely optimal sequence (Akashi 1995).

1.5.1 Synonymous Substitution Rates

Perhaps the most fruitful approach to gaining insight into the process of molecular evolution, and a useful means of gauging the functional significance of sites within sequences, is the comparison of homologous sequences between closely related species (Li and Graur 1991). Silent substitutions normally occur at a much higher frequency than non-synonymous substitutions (Kimura 1977; Li *et al*. 1985b). As the process of substitution is readily identifiable, synonymous mutations have been subjected to intense study because they have the potential to reveal many of the forces that underlie molecular evolution (Sharp *et al*. 1995).

The rate of evolution at synonymous sites has been used to investigate and validate some of the predictions of molecular evolution, such as the molecular clock hypothesis (Fitch and Strausbaugh 1993; Morton 1994; Wolfe, Sharp and Li 1989). The rate of silent substitutions is substantially lower in highly expressed genes than in genes expressed at lower expression levels (Ikemura 1985; Sharp 1991; Sharp and Li 1987b). The observation that constraint on codon usage reduces silent substitution rates and that this constraint can vary between genes, is consistent with the predictions of the neutral theory (Kimura 1983). Synonymous substitutions vary at a number of different scales of resolution, between genomes, across a single genome and within genes. Near the initiation codon of *E. coli* genes the rate of synonymous substitution is lower, suggesting additional selection pressure in this region (Eyre-Walker and Bulmer 1993).

Synonymous substitutions can elevate the rate of substitution in an adjacent codon by about 10%; this appears to be unrelated to the level of gene expression and has a small range of influence. This may be due to sequence directed mutagenesis, recombination and/or selection (Eyre-Walker 1994c). Neighbour mutation bias was estimated in *E. coli* and yeast, where a similar pattern was found in complementary sequences in the synonymous usage of A vs. G and U vs. C. This reflected a codon context effect on mutation patterns in weakly expressed genes (Bulmer 1990). Wide variation in neighbour substitution rates have also been found in other species, where again the nearest neighbour base can influence the substitution rates (Blake *et al*. 1992).

The relationship between codon usage and the rate of substitution at silent sites is more complex than just selection for optimal codons. While the increase of expression increases the selection pressure on synonymous codons and directly reduces observed substitution rates, there is also a decrease in the mutation rate (Berg and Martelius 1995). In *E. coli* this decline in mutation rate appears similar for the lysine family of codons which do not appear to be strongly selected for translational efficiency, and for phenylalanine codons, which are selected for translational optimality (Eyre-Walker and Bulmer 1995).

Among the eukaryotes synonymous substitutions have been most extensively studied in mammals, where significant variations in K_s have been found (Li *et al*. 1985a). Synonymous substitution rates in mammals are gene specific and correlated with frequencies of non-synonymous substitutions. Silent site substitution rates between human and murid (mouse and rat) genes are similar for neighbouring genes but vary around the genome (Matassi *et al*. 1999). If synonymous substitutions are indeed essentially neutral it implies that mutation rates are varying systematically (Sharp *et al*. 1995). This is most easily explained when the presence of isochores is considered; presumably, the different isochore types have different local mutation biases/rates. This in turn may explain why the molecular clocks of mammalian genes differ (Wolfe, Sharp and Li 1989).

By comparing polymorphism and divergence in *Drosophila* between putatively favourable and deleterious codons, it was shown that even weak selection could substantially alter ratios of polymorphism to divergence from that expected under neutrality (Akashi 1995).

1.6 Does Codon Usage Regulate Expression?

Rare codons can be defined based on the overall codon usage, or the codon usage of highly biased genes (Kane 1995; Zhang *et al*. 1991). Rare *E. coli* codons include AGG (Arg), AGA (Arg), AUA (Ile), CGA (Arg), CUA (Leu) and GGA (Gly) (Grosjean and Fiers 1982; Sharp *et al*. 1988). The choice of low usage codons is relatively insensitive to gross base composition, with some codons (e.g. CGG) relatively infrequent in a wide range of species including *E. coli*, *Drosophila*, primates and yeast (Zhang, Zubay and Goldman 1991).

The frequency of rare codons is higher in rarely transcribed genes. Often this is ascribed to adaptive pressures modulating gene expression. The low level of expression of *dnaG*, which is cotranscribed with the highly expressed *rpsU* and *rpoD* genes, was attributed to its higher frequency of rare codons, even though it was noted that *dnaG* had a weak ribosomal binding site (Konigsberg and Godson 1983). Models where the rate of polypeptide elongation is regulated by the presence of rare codons are frequently invoked by molecular biologists to explain their presence. Generally these models suppose that stabilising selection operates to maintain a certain level of codon usage bias. The models, which have been described by Kimura (1983) as pan-selectionist codon usage models, have gained wide acceptance. Much of the experimental literature on codon bias appears to be devoted to what is accepted as the self evident proposition that rare codons regulate gene expression by regulating translation rates (Grosjean and Fiers 1982; Hoekema *et al*. 1987; Konigsberg and Godson 1983; Robinson *et al*. 1984; Varenne *et al*. 1984). Population geneticists frequently challenge these models however, by arguing that the presence of rare codons is due to drift randomising codon usage (Bulmer 1991; Holm 1986; Ikemura 1985; Kurland 1993; Li 1987; Sharp and Li 1986; Shields 1989).

There is supporting evidence for the hypothesis that the higher frequency of rare codons in lowly expressed genes reflects mutation biases rather than positive selection for rare codons. Indeed it is not obvious how pan-selectionist models can explain the observed uniform patterns of codon usage (Sharp and Cowe 1991). Lowly biased genes display other influences of mutation bias; e.g., codon contexts are strongly influenced by neighbouring bases. The frequencies of dinucleotides and their complimentary dinucleotides are similar (Bulmer 1990). Rare codons in genes with low expression levels are not under strong selective pressures (Sharp and Li 1986). Substitutions accumulate as quickly in the regulatory genes, *dnaG* and *araC*, as in other lowly biased genes (Sharp and Li 1987b). Rather than being positively selected in lowly expressed genes, rare codons are under a strong negative selection in highly expressed genes. The level of expression determines rare codon usage and not *vice versa*.

Although many experimental results apparently supported the hypothesis that the presence of rare codons directly effects yield of product, their interpretation may be overly simplistic. For instance, Ivanov *et al.* (1992) demonstrated that the rare AGG doublet, which had been reported to have an inhibitory effect in *E. coli* (Robinson *et al*. 1984), had an equally inhibitory effect whether located 5’ or 3’ of the start codon. Similarly, Brown (1989) observed that part of the *pgk* gene mutatgenized by Hoekema *et al.* (1987) contained a transcriptional activator. It is also evident that small changes in the primary mRNA structure can have large effects on mRNA stability (Petersen 1987). Few of the papers on the effect of codon usage on expression level take into account any changes in mRNA half-life (Kurland 1991).

In principle, strings of rare codons could synergistically increase translation time, but not translation rate unless they affect the rate of ribosomal binding (Sorensen, Kurland and Pedersen 1989). The insertion of nine consecutive low-usage CUA (Leu) codons immediately downstream of codon 13 of a 313-codon test mRNA strongly inhibited its translation without apparent effect on translation of other mRNAs containing CUA codons (Goldman *et al*. 1995). In contrast, nine consecutive high-usage CUG (Leu) codons at the same position had no apparent effect, and neither low nor high-usage codons affected translation when inserted after codons 223 or 307. The strong positional effects of the low-usage codons could not be explained by differences in stability of the mRNAs or in stringency of selection of the correct tRNA. It could be explained by translation complexes being less stable near the beginning of a message, slow translation through low usage codons early in the message might cause translation complexes to dissociate before completing the read through (Goldman *et al*. 1995). The rare UUA codon only affected product yields when located near the start codon (Goldman *et al*. 1995). The inhibitory effect was reduced when positioned more than 50 codons from the initiation codon, or by overexpression of the *argU* gene (tRNA$_{\text{arg}}$ UCU/CCU) (Chen and Inouye 1994). This has been interpreted as evidence that the increased frequency of less commonly used codons near the start of genes plays an important role in the regulation of gene expression (Chen and Inouye 1990; Chen and Inouye 1994).

Some proteins that contain a high percentage of low usage codons have been described as belonging to families where an excess of the protein could be detrimental to fitness (Zhang, Zubay and Goldman 1991). Saier (1995) has discussed how the inappropriate expression of certain genes might be globally regulated by altering the pool of tRNAs at different stages of growth. For example, the codon usage of genes encoding the photosynthetic apparatus of the Gram-negative purple bacterium *Rhodobacter spheroides* differs from genes encoding the fructose pathway (Wu and Saier 1991). This may in part be due to different tRNA pools under photosynthetic growth relative to heterotrophic growth (Saier 1995). In *Clostridium acetobutylicum* a mutation in the *thrA* genes (tRNA$_{\text{thr}}$ ACG) causes loss of solventgenesis, the codon ACG is rarely used and largely restricted to genes expressed after exponential growth (Saier 1995).

Streptomyces species can enter a vegetative growth phase, during which they can produce antibiotics and other useful secondary metabolites. Mutations, including deletions, of the *Streptomyces coelicolor bldA* gene (tRNA$_{\text{leu}}$ UUA) prevent efficient phenotypic expression of several genes that are normally expressed during vegetative growth and which contain the rare leucine codon UUA (Ueda *et al*. 1993). In wild type cells tRNA$_{\text{leu}}$ UUA accumulates in ever-increasing amounts as *S. coelicolor* ages. The deletion mutations of *bldA* did not prevent vegetative growth but stopped mycelium formation and the production of secondary metabolites. The presence of UUA codons in recombinant proteins also inhibits foreign gene expression in *Streptomyces lividans* (Ueda *et al*. 1993).

Again, the interpretation of these results is difficult. While there is a higher frequency of rare codons near the initiation codon of many regulatory genes there is also a higher frequency of rare codons near the initiation codons of highly expressed genes (Eyre-Walker and Bulmer 1993). The differentiation of codon usage patterns at different stages of growth is not necessarily a regulatory mechanism. It may simply reflect the difference in the mechanisms controlling tRNA abundance during exponential and stationary growth phase (discussed above) and a consequent adaptation to different tRNA pools.

An early investigation involved the addition of four rare AGG (Arg) codons near the initiation codon of a reporter gene. The yield of product was compared with a control gene that contained four common CGT (Arg) codons at the same positions (Robinson *et al*. 1984). Under conditions of maximum expression levels at least one third less protein was synthesised by constructs containing the rare codons, but at lower levels of expression the constructs produced a similar yield of products (Robinson *et al*. 1984).

1.6.1 Programmed Frame Shifting

Recoding is the term given to programmed alteration in the reading of the genetic code (Gesteland *et al*. 1992), and is observed in a minority of sequences in probably all organisms (Larsen *et al*. 1996). Where recoding occurs there are often sites associated with elevating the frequency of recoding. The majority of these sites are 3’ to the shift site, though there have been several 5’ stimulators found. The first was found in the *prfB* gene, which encodes release factor 2 (RF-2). The RF-2 protein mediates polypeptide chain release at UGA and UAA codons. The expression of RF-2 is autoregulated (Craigen *et al*. 1985) the zero frame of the protein has a stop codon UGA at the 25th codon. If RF-2 is limiting, the ribosome will +1 frameshift to allow expression. This phenomenon is also exploited as an assay system for the measurement of codon recognition and accuracy (Curran 1995).

The minimal sequence of *prfB* mRNA necessary for efficient +1 frameshifting includes the frameshift site and an additional crucial Shine Dalgarno (SD) like element (Weiss *et al*. 1987). Located three bases 5’ of the CUU shift codon, this SD sequence (AGGAGG) is not involved in translational initiation, but pairs with the 3’ end of the elongating ribosome (Weiss *et al*. 1988). The spacing between this SD sequence and the shift site is critical to the frameshifting (Weiss *et al*. 1987). It seems reasonable to infer that the SD interaction acts to stimulate frameshifting by decreasing termination (Larsen *et al*. 1996).

The translation of the AGG doublet can result in a 50% frame shift (Spanjaard and van Duin 1988). The insertion of between two and five AGG codons six codons prior to the termination codon, at high expression levels, increases the production of aberrant proteins without affecting mRNA stability (Rosenberg *et al*. 1993). The yield of aberrant product increases as the number of AGG codons increases, this is consistent with the hypothesis that at sufficiently high concentrations of AGG-containing mRNA, all the tRNA $_{\text{AGG}}$ is sequestered. Thus translation stalls at the AGG codons stimulating frameshift, hop or termination (Rosenberg *et al*. 1993).

1.6.2 Rare Codon Usage may be Correlated with Pause Sites

Besides affecting the overall rate of translation, synonym choice may be involved in influencing fluctuations in the elongation rate along the mRNA. It has been suggested that rare codons may be clustered to facilitate ribosomal pausing at sites corresponding to protein domain boundaries (McNally *et al*. 1989; Purvis *et al*. 1987). This hypothesis was presented by Purvis *et al.* (1987) was based on the observation of an apparent cluster of rare codons in the *S. cerevisiae pyk* gene. This region was later resequenced and was found to be a sequencing artefact, though the authors still felt that their theory was still tenable (McNally *et al*. 1989). It has also been proposed that translational pausing could favour protein export by increasing the time required for translation elongation, thus allowing time for nascent polypeptide to be exposed to the cytoplasm and facilitate chaperone binding. However the distribution of rare codons is independent of polypeptide length and thus does not seem to support the export theory (Collins *et al*. 1995).

1.6.3 Codon usage and heterologous gene expression

E. coli remains a popular choice for the expression of heterologous proteins. The presence of rare codons *per se* does not imply weak expression. Despite the poor overlap between the codon usage of *Halobacterium halobium* (70% G+C) and *E. coli* (50% G+C), genes from *H. halobium* can be highly expressed in *E. coli* (Nassal *et al*. 1987). Similarly the *pepC* gene from *Lactobacillus delbrueckii* ssp. *lactis* can be over expressed in *E. coli* (Klein *et al*. 1994). In *E. coli* mutation of the ribosomal binding site of *atpH* can increase its level of expression 20-fold (Rex *et al*. 1994). An oligonucleotide of rare codons within the coding sequence of *B. subtilis sspB* (small acid soluble spore-protein) did not have a discernible effect on yield (Loshon *et al*. 1989). The addition of rare AGG codons near the terminus actually enhanced expression of chloramphenicol acetyltransferase in *E. coli* (Gursky and Beabealashvilli 1994). The frequency of rare *E. coli* codons in protozoan parasites had been predicted to have implications for their expression in *E. coli* (Sayers *et al*. 1995). Despite this, expression of *Trypanosoma* genes is up to 20 fold higher in *E. coli* then in their natural genome (Isacchi *et al*. 1993).

However, the expression of heterologous genes can be adversely affected by unusual codon usage or context (Kane 1995). For example, the expression of bovine placental lactogen in *E. coli* results in a 2 codon frameshift (Kane *et al*. 1993) and the expression of human transferrin in *E. coli* results in 2% to 4% +1 frameshifting, at a CCC-UGA site (de Smit *et al*. 1994). The presence of rare codons in a recombinant gene can be compensated for by either adding the appropriate tRNA, or synthesising the gene to remove the rare codons. The expression in *E. coli* of the human granulocyte macrophage stimulating factor was enhanced after *argU* was induced (even though the recombinant protein had only a single AGG codon) (Hua *et al*. 1994). The human *rap74* gene (RNA polymerase associating protein) was expressed more efficiently in *E. coli* after codon usage was adjusted, previously there are a large number of amino terminal fragments due to frameshifts (Wang *et al*. 1994). Similarly altering the codon usage of avidin (Airenne *et al*. 1994), tropoelastin (Martin *et al*. 1995) and isovaleryl-coa dehydrogenase (Mohsen and Vockley 1995) enhanced their expression in *E. coli*.

The influence of codon usage on gene expression has also been used as a rationale for the choice of recombinant host. Based on the similarity of codon usage *Bacillus thuringiensis* was recommended as a recombinant host for expressing plant genes from *Brassica* (Kumar and

Sharma 1995). The codon usage patterns and ribosomal binding sites of highly expressed cyanobacterial genes, suggested that the cyanobacterium *Synechococcus* pcc-7942 would be an inappropriate host for the expression of the larvicidal *B. thuringiensis cryiVB* gene (Soltesrak *et al*. 1995).

1.7 Codon Usage as a Tool for Gene Prediction

Knowledge of codon usage preference can be applied to the prediction of open reading frames (Borodovsky *et al*. 1995; Krogh *et al*. 1994; Staden and Mclachlan 1982). With the arrival of the large scale sequencing projects, the prediction of gene introns and exons has become of paramount interest. Most of the many modern gene prediction programmes use codon usage patterns as well as dinucleotide and short oligonucleotide patterns to predict open reading frames (Karlin and Cardon 1994). The GeneMark prediction programme (Borodovsky *et al*. 1994a; Borodovsky and McIninch 1993; Borodovsky *et al*. 1994b) has been used to identify the coding sequences from two major shotgun genome sequencing projects (Fleischmann *et al*. 1995; Fraser *et al*. 1995).

Although modern gene prediction programs can learn from a sample of genes, a more in depth knowledge of codon usage variations can greatly improve their predictive properties (Borodovsky *et al*. 1995). Applying GeneMark to the prediction of genes in *E. coli*, found that the detection of class III genes (Medigue *et al*. 1991) was the most difficult and that they were easily overlooked by inappropriate parameters. Class III genes could only be identified by GeneMark with any degree of accuracy if the programme was trained on a representative sample of class III genes, unlike class I and class II genes which can be recognised with a low error rate when trained on either set (Borodovsky *et al*. 1995).

1.8 Analysis of Codon Usage

Compilations of codon usage are of limited value due to the complexity of the information. Too often, the tabulation of codon usage is the only codon usage analysis presented, even when there is enough data to generate an in-depth analysis of codon usage variation (Forsburg 1994; Wada *et al*. 1991; Wada *et al*. 1992; Winkler and Wood 1988). Early analysis of codon usage pooled the codon usage from different sets of genes and then calculated and compared the biases (Berger 1978). Such analyses required either the *a priori* grouping of genes or a prohibitive number of pair wise comparisons. The significance of such tests was strongly influenced by sample size and was dependent on the assumptions used for the groupings. As the number of sequenced genes increased this type of analysis became impractical. A major advance in the analyses of codon usage was pioneered by Grantham and co-workers, when they applied multivariate statistical techniques to the investigation of codon usage (Grantham, Gautier and Gouy 1980a; Grantham *et al*. 1980b). A second major advance was the application of simple indices that could summarise optimal codon usage into useful descriptive variables, facilitating the comparison of codon usage patterns (Bennetzen and Hall 1982; Gouy and Gautier 1982).

1.8.1 Multivariate Analysis

The purpose of statistics has been described as to summarise, simplify and eventually explain (Greenacre 1984). Codon usage by its very nature is multivariate, it is therefore necessary to analysis this data with multivariate statistical techniques. If one examines the set of conventional statistical techniques in use today, it is clear that the statistician can rarely proceed without introducing a certain degree of subjectivity into the analysis (Greenacre 1984). It is therefore advantageous to use a method that can examine data without *a priori* assumptions and there are several multivariate analysis techniques that satisfy this condition (Greenacre 1984).

Multivariate analyses (MVA) are used to simplify rectangular matrices in which (for these purposes) the columns represent some measurement of codon or amino acid usage and the rows represent individual genes. Examples of MVA techniques that have been successfully applied to the analysis of codon usage are cluster analysis and correspondence analysis. Cluster analysis partitions data into discrete groups based on the trends within the data, but has the disadvantage that it sometimes forces arbitrary divisions of a dataset even when presented with continuous variation. Correspondence analysis is an ordination technique that identifies the major trends in the variation of the data and distributes genes along continuous axes in accordance with these trends. Correspondence analysis has the advantage that it does not assume that the data is in discrete clusters and can represent continuous variation accurately. It is also possible to combine both methods by superimposing cluster analysis onto correspondence analysis (Grantham *et al*. 1981; Grantham *et al*. 1980b; Medigue *et al*. 1991). Codon usage has been analysed using other MVA methods, these include multidimensional scaling ordination and eigenanalysis ordination (Morrison, Ellis and Johnson 1994). Graphical methods can also be useful in displaying multivariate data (Wainer 1983) and has been used to analyse codon usage in *E. coli* (Zhang and Chou 1994). Multivariate analysis has also been applied to other biological questions, including the prediction of coding sequences (Fichant and Gautier 1987), the analysis of phylogeny (Higgins 1992), and the analysis of the amino acid usage patterns among genes (Lobry and Gautier 1994).

1.8.1.1 Correspondence Analysis

Correspondence analysis (COA) has been reinvented many times, in many disciplines. Ecologists refer to correspondence analysis as “reciprocal averaging” while psychologists refer to it as “dual optimal scaling” (Greenacre 1984). This reinvention is possible because the theory of correspondence analysis can be, and has been, derived in many different ways.

The name correspondence analysis has lost something in the translation from the French “Analyse des correspondances”, literally the “analysis of correspondences” where the French term “correspondance” is used to denote the “system of associations” between rows and columns (Greenacre 1984). The technique was unfavoured in the Anglo school of statisticians during the 1960s (Greenacre 1984), so much so, that when Hill (1974) tried to revitalise it he described it as a “neglected technique”. However correspondence analysis was still popular among the French school of analysis (Benzecri 1992), where it was primarily developed for the analysis of language.

The French geometric approach to correspondence analysis was initiated by Benzecri around 1965 (Greenacre 1984). The group lead by Benzecri has a philosophy of data analysis founded on inductive reasoning. One of their primary philosophies was that it was the dataset which was important, not the model, which must fit the data and not *vice versa* (Greenacre 1984).

For a complete mathematical description of correspondence analysis, the reader should see texts by Greenacre (1984), Lebart *et al.* (1984) and the translation of the work by Benzecri (1992). In simplified form, each gene can be represented in multidimensional space by a vector. The multidimensional space is Euclidean, with each axis orthogonal. Distances are defined in rows and/or columns of a table, and these distances are approximated by Euclidean distances in a low-dimensional representation of the table. Thus COA is a form of metric multidimensional scaling (Benzecri 1992). The objective of correspondence analysis is to identify a lower dimensional subspace that best represents the data points. As correspondence analysis treats the rows and columns of the data matrix in a symmetric fashion (Lebart *et al*. 1984), and so allows the rows and columns to be represented in the same lower dimensional space. For algebraic simplicity, the goodness of fit is estimated on squared distances. Points can be weighted by an associated mass; points with zero effect on the lower subspace have masses of zero. (Greenacre 1984).

Correspondence analysis of codon usage is an exploratory graphical approach that does not test the significance of associations between genes and codons, but merely identifies such associations. It is a powerful tool to isolate trends, but its major weakness is that it provides no clues as to the interpretation of those trends. Correlations are difficult to judge and it is not necessarily legitimate to equate correlation with causation. Correspondence analysis of codon usage can distinguish highly and lowly expressed genes (Gouy and Gautier 1982; Grantham *et al*. 1981; Holm 1986) and is the most widely used multivariate technique for the analysis of codon usage.

Principal component analysis (PCA), a related MVA technique, differs from correspondence analysis in one important respect. In PCA, the columns of the data matrix are generally a set of measurements or variables whereas the rows are a relatively homogeneous sample of objects. PCA analysis can be considered a theoretically appropriate method for the analysis of data that derives from a multivariate normal distribution. COA is the more theoretically appropriate method for the measurement of data which is in the form of a contingency table (Lebart, Morineau and Warwick 1984). Principal component analysis can be used for the differentiation of genes according to their codon usage but for historical reasons, the method of choice remains COA.

1.8.1.2 Cluster Analysis

Cluster analysis has been described “as the art of finding groups in data” (Kaufman and Rousseeuw 1990). There are two main types of clustering algorithms namely partitioning and hierarchical methods. Under the partitioning method, each group must contain at least one object and each object must belong to at least one group. The objective of partitioning clusters is to minimise the average distance between objects within a group, and to maximise the average distance between the groups. If the method produces k clusters from n objects then $k \leq n$. It is important to note that k is user defined, although it can be chosen automatically. An inappropriate k value can produce an “unnatural” clustering.

In genetics, the most familiar forms of cluster analysis are the hierarchical methods such as unweighted pair group mean average method (UPGMA) and the neighbour joining method (NJ) in the construction of phylogeny. Hierarchical methods can be either agglomerative or divisive, both NJ and UPGMA are agglomerative. Divisive methods are not commonly employed for large datasets because they must consider all possible divisions of the data into two subsets, a process which is very computationally intensive (Kaufman and Rousseeuw 1990).

Partitioning and hierarchical cluster analyses have both been applied to the analyses of codon usage. Hierarchical cluster analysis of the relative synonymous codon usage (RSCU) usage of 100 *S. cerevisiae* genes, found apparently bimodal codon usage. The dendrogram defined two groups of genes, with those genes that might expect to be highly expressed located within the same cluster (Sharp, Tuohy and Mosurski 1986). Thus, cluster analysis differentiated highly and lowly expressed genes. Cluster analysis is somewhat dependent on knowing the patterning of clustering underlying the data. If the yeast genes had not fallen into clearly defined groups, the pattern of codon usage would have been much less clear-cut. Cluster analysis was used be Medigue et al. (1991) to group 780 *E. coli* genes by codon usage into three defined classes.

1.8.2 Indices

Codon usage indices are used to help the tabulation and investigation of codon usage. Indices can reduce the codon usage data into a useful summary, but can have certain limiting assumptions. There are two basic types of codon usage indices. One measures the overall deviation of codon usage from some expected usage and the other measures a bias towards a particular subset of optimal or preferred codons.

1.8.2.1 Indices of Codon Usage Deviation.

These indices measure the deviation of observed codon usage from some expected codon usage distribution. The two main null hypotheses used to estimate expected codon usage distributions are:

1 H_0: The expected codon usage is entirely determined by mutation bias.

2 H_0*: A special case of H_0, which assumes no mutation bias, i.e. codons are used equally.

H_0* is the most commonly used null hypothesis because it makes the simplest assumptions. However, equal usage of synonyms is the exception rather than the rule. Indices such as chi based methods (G-statistic (Sharp, Tuohy and Mosurski 1986) and scaled chi squared (Shields *et al*. 1988), can be used with reference patterns other than H_0*. Other indices such as the effective number of codons (N_c) (Wright 1990), although most often used under H_0*, are equally applicable under H_0.

1.8.2.1.1 P2

The P2 index (Gouy and Gautier 1982) is one of the earliest codon usage indices, it calculates the proportion of codons that conform to the intermediate strength of codon-anticodon interaction rule (Grosjean and Fiers 1982).

$P2 = \frac{(WWC + SSU)}{(WWY + SSY)}$ Where W= A or U; S= G or C; Y= C or U.

This index is not independent of rare codon usage. A modification of this index P2′ which excludes CCY (CCC is rare in *E. coli*) has been used to quantify *E. coli* codon usage (Sharp and Li 1986). Under uniform codon usage (H_0*) P2 is equal to 0.5.

1.8.2.1.2 P

The codon preference plot (P) (Gribskov *et al*. 1984) index is calculated as follows:

$$p = \frac{\left(\frac{f_x}{F_x}\right)}{\left(\frac{r_x}{R_x}\right)}$$

where f_x is the frequency of codon_x; F_x is the frequency of the synonymous family of codons; r_x and R_x random usage of these codons, assuming the same base composition as the sequence of interest. P can be considered as the relative likelihood of a codon being found in a gene, as opposed to being found in a random DNA sequence.

1.8.2.1.3 GC3s

The index GC_{3s}, is the frequency of G or C nucleotides present at the third position of synonymous codons (i.e. excluding Met, Trp and termination codons).

1.8.2.1.4 GC skew

Measures the skew in the frequency of G and C nucleotides

$$GCskew = \frac{G-C}{G+C}$$

1.8.2.1.4.1.1.1 ENc

The effective number of codons (ENc) used by a gene (Wright 1990) is a simple measure of codon bias. It is analogous to the effective number of alleles used in population genetics. Its value represents the number of equally used codons that would generate the same codon usage bias as that observed. It can be calculated without reference to optimal codons. It is a general measure of bias away from equal usage of alternative synonyms (i.e. it is a measurement of the 'true' number of codons used). In an extremely biased gene where only one codon is used for each amino acid this would be 20 and in an unbiased gene (assuming H_0*) it would be 61. In genomes where codon usage is entirely due to mutational bias (H_0) the expected value of N_c (depending on the degree of G+C bias) ranges from to 31 and 61 (Wright 1990).

1.8.2.1.4.1.1.2 RSCU

Relative synonymous codon usage (RSCU) (Sharp, Tuohy and Mosurski 1986) is calculated as the ratio of the observed frequency of a codon to the frequency expected if codon usage was uniform (H_0*) within a synonymous codon group (Hastings and Emerson 1983). RSCU values close to 1.0 indicate a lack of codon bias. RSCU values are largely independent of amino acid composition and are particularly useful in comparing codon usage among genes, or sets of genes that differ in their size and amino acid composition.

1.8.2.1.4.1.1.3 χ^2/n and G statistic

Scaled chi squared indices have been used to measure bias between observed and expected codon usage, with uniform usage (H_0*) as the expected codon usage (Shields *et al*. 1988). The chi squared statistical values were corrected using Yates correction for small sample sizes and scaled by dividing the chi value by the number of synonymous codons (n). The scaled G statistic is similar to the scaled chi squared index. It also compares the observed frequency of synonymous codons with those expected under random codon usage (H_0*) (Sharp and Li 1986) and is scaled by the number of synonymous codons. Neither the G statistic nor the scaled chi-square, are true tests of significance; rather they are indices of bias. Neither index performs well if the gene is very short.

1.8.2.2 Indices that Measure Codon Bias towards a subset of Preferred Codons.

In many species, highly expressed genes preferentially use a subset of codons (i.e. optimum codons). Several indices estimate the extent to which the codon usage of a gene has been altered towards the preferential usage of these optimal codons.

1.8.2.2.1 CBI

Codon bias index (CBI) is a measure of directional codon bias towards a subset of optimal codons (Bennetzen and Hall 1982).

$$CBI = \frac{Nopt - Nran}{Ntot - Nran}$$

Where N_{opt} = number of optimal codons; N_{tot} = number of synonymous codons; N_{ran} = expected number of optimal codons if codons were assigned randomly (H_0*). CBI is similar to F_{op} (Ikemura 1985), except that N_{ran} is used as a scaling factor. In a gene with extreme codon bias, CBI may equal 1.0. With random codon usage CBI would be zero. If N_{opt} is less than N_{ran}, the index will be negative.

1.8.2.2.2 F_{OP}

This index measures the frequency of optimum codons (F_{op}) in a gene (Ikemura 1981a; Ikemura 1985; Ikemura and Ozeki 1982). It is a species-specific measure of bias towards particular codons that appear to be translationally optimal in a species. It is a simple ratio between the frequency of optimal codons and the total number of synonymous codons. Its values range from 0 (when a gene contains no optimal codons) to 1 (when a gene is entirely composed of optimal codons).

1.8.2.2.3 CAI

The codon adaptation index (CAI) (Sharp and Li 1987a) is a very widely used measure of codon bias in prokaryotes (Eyre-Walker and Bulmer 1993; Gutierrez *et al*. 1994; Perriere *et al*. 1994) and eukaryotes (Akashi 1994; Frohlich and Wells 1994; Morton 1994). It summarises the adaptation of codon usage towards the codon usage of a set of genes known to be highly expressed. The frequency of codon usage in the highly expressed genes is used to define relative fitness values for each synonymous codon. These values are calculated based on RSCU rather than raw codon usage and are therefore essentially independent of amino acid composition.

CAI avoids the dichotomy inherent in both the F_{op} and CBI indices where codons are either optimal or non-optimal, thus small changes in sample size can have the undesirable effect of changing the subset of optimal codons. However, the CAI values of genes from different species are not directly comparable, because the relative fitness values differ. Similarly if the reference set of highly expressed genes changes, the relative fitness values will change and the CAI values for all genes from that species must be recalculated.

The CBI, $F_{op,}$ and CAI indices measure bias towards preferred codons. They should only be calculated for those species where selection for the translational efficiency has overcome mutational drift and a set of optimal codons (or a set of highly expressed genes) has been identified. If these indices are calculated for genes where the optimal codons are unknown or where codon usage is determined by mutational bias, the resulting index value is essentially meaningless. For example, the CAI values of *H. sapiens* genes have no “meaning” because human codon usage is driven by mutational biases (Sharp *et al*. 1993) but this has not prevented the CAI value of human genes being reported (Brown *et al*. 1994).

1.9 Proposal

It is widely accepted that non-random codon usage can be influenced by both natural selection and mutational biases. The strength and direction of these forces vary within and between genomes. These forces have led to a considerable heterogeneity in codon usage patterns among different genes and species. Despite an ever-increasing literature on the codon usage, there were relatively few reports where codon usage had been analysed to discover if the observed biases in codon usage were due to mutational biases, selection or a combination of both. Many reports, consisted of simple tabulations of codon usage and made the assumption that a gene with a high codon usage bias would be highly expressed. There were few reports of any form of MVA of codon usage (the technique of choice). There was also little consistency in the types or the methodologies of the analyses performed, which made comparisons between reports quite difficult. Many of these observations could be attributed to a scarcity of software designed for the analysis of codon usage and in particular the MVA analysis of codon usage.

Historically the group of Prof. P. M. Sharp used a mixture of programmes from several sources to analysis codon usage. The correspondence analysis software was a FORTRAN 77 programme DECORANA written by M. O. Hill, which while invaluable had only a limited functionality. For example, it was not possible to store or reuse the vectors identified during the analysis of one dataset on another, only the first four factors were reported, and the relative inertia that each gene or codon explained of each factor (and *vice versa*) was not calculated.

Therefore, one of the main objectives of this study was to develop a software package for the analysis of codon usage variation. This package, CodonW, was to be in the public domain, portable, but also powerful and capable of performing quite sophisticated codon usage analyses. It was designed to have both a menu and a command line interface and perform all the analyses required for an in-depth investigation of codon usage. This ranges from the tabulation of overall codon usage right through to correspondence analysis. It was designed to be flexible enough to analyse both codons and amino acid usage, and to generate its results in an easily comprehensible format. A description of CodonW and examples of its application to the analysis of codon usage and amino acid usage are presented in Chapter 2.

CodonW was also used to investigate the codon and amino acid usage in species where our understanding of codon usage remained fragmentary. During the period of this thesis the sequence of the whole genome of cyanobacterium *Synechocystis* PCC 6803 was published, a previous analysis of the codon usage of a subset of cyanobacterium *Synechocystis* PCC 6803 found no detectable difference between the codon usage of putative highly and lowly expressed genes (Krishnaswamy and Shanmugasundaram 1995; Malakhov and Semenenko 1994). This is interesting, as it would imply that in this species natural selection had little effect on shaping codon usage, despite the genome not having a particularly biased base composition. An analysis of the codon usage of *Synechocystis* PCC 6803 is presented in this thesis.

An understanding of codon usage of the low G+C Gram positives was effectively limited to that of *B. subtilis*. There had however been several reports of apparent bias in the codon usage of the low G+C Gram positive *Lactococcus lactis*, and in the codon usage of its phosphoenol pyruvate transport (PE-PTS*) lac* operon. An analysis of the codon usage of *L. lactis* and this operon is presented in this thesis.

The PE-PTS *lac* operon has been reported to be present in the low G+C Gram positive species *Staphylococcus aureus* and *Streptococcus mutans*. A comparative analysis of the codon usage of these species and their *lac* operons with each other, and with *L. lactis* is also presented.

2 CodonW

2.1 Introduction

Codon usage analysis has a number of practical applications, one of which is the inference of expression level. The most highly expressed genes in *E. coli, S. cerevisiae* and *B. subtilis* are also those with the strongest codon bias (Sharp and Cowe 1991; Sharp and Li 1987a; Shields and Sharp 1987) and this has been used to infer that genes with high codon bias are highly expressed (Cancilla *et al*. 1995b; Freirepicos *et al*. 1994; Gharbia *et al*. 1995). The reciprocal inference, that genes with low codon bias are lowly expressed genes, does not always hold true (Fitch and Strausbaugh 1993; Kliman and Hey 1993; Kliman and Hey 1994).

The limiting factor preventing a rigorous analysis of codon usage is often lack of sequence information. A prerequisite of any analysis is that the sample size is large enough. This is dependent upon the magnitude of the selective coefficient, the effective population size and the degree of overlap between the translationally advantageous codons and those codons generated by mutation biases. These factors are often unknown before analysis is performed, and must be established empirically. Hence the lack of evidence for positive selection of translationally optimal codons is not necessarily evidence for the absence of selection, as exemplified in *B. subtilis* (see introduction) (Ogasawara 1985; Shields and Sharp 1987). The choice of protein genes to be included in any analysis is important. They should be representative of the codon usage of the host genome, this excludes genes “recently” acquired by horizontally transfer (e.g. many genes associated with antibiotic resistance, plasmids, phage and transposable elements). They should include (or at least be expected to include) a mixture of high and low expression levels, as optimal codons are usually identified by comparing and contrasting the codon usage of (putatively) highly and lowly expressed genes. To limit variation due to stochastic noise, short genes (less than 50 codons) are normally excluded. Even with these restrictions there are many species whose genomes sequences have yet to be published (e.g. *Clostridium acetobutylicum, Streptococcus mutans, Staphylococcus aureus,* cyanobacteria *Synechococcus sp, Schizosaccharomyces pombe* and *Lactococcus lactis*) where the quantity, expression level and origin of sequenced genes make a detailed examination of codon usage feasible, yet in which only a preliminary examination has been reported (Cancilla *et al*. 1995a; Croux and

Garcia 1992; Forsburg 1994; Krishnaswamy and Shanmugasundaram 1995; Popplewell *et al*. 1991; Wada *et al*. 1992).

Despite the interest in codon usage, the proportion of publications that include a statistical analysis of codon usage remains low. This may be due to the (perhaps perceived) lack of programs for codon usage analysis. The analytical methods used range from the tabulation of codon usage to multivariate statistical analysis (MVA). For many workers the more sophisticated statistical techniques that have been used to analyses codon usage (e.g. correspondence analysis) are not readily available or accessible. Previously COA has often required access to specialised software such as DECORANA (written by M. O. Hill), CORAN (Lebart, Morineau and Warwick 1984) or commercial statistical packages such as SAS or SPSS. These programs or packages are general statistical packages and are not designed specifically to deal with biological problems. Without specialised software it is not a trivial task to convert sequence information (as represented by a linear DNA or RNA strand) into the numerical input formats required by these statistical packages. Two programs, MacMul and NetMul, have been developed to simplify MVA of biological, but not specifically sequence data (Thioulouse 1990a; Thioulouse 1990b; Thioulouse and Chevenet 1996; Thioulouse *et al.* 1995).

2.1.1 Codon Usage Analysis software

There is a surprising scarcity of software designed to aid the analysis of codon usage. There are programs available that tabulate species codon usage directly from the public nucleic acid databases (GenBank) (Nakamura *et al.* 1996). Some programs can calculate particular codon usage indices such as: the effective number of codons (N_c) (Wright 1990); the codon bias index (CBI) (Bennetzen and Hall 1982); or more general indices of codon bias (Goldman *et al*. 1995; Krishnaswamy and Shanmugasundaram 1995; Rodriguezbelmonte *et al*. 1996). However these programs use different input formats and restrict themselves to the calculation of a single index. One program that has integrated several codon usage indices with other measures of codon composition was CODONS (Lloyd and Sharp 1992), written in FORTRAN 77 (ANSI 1978). CODONS served as the inspiration for the package described herein.

GCG (GCG 1994), the most commonly used sequence analysis package includes programs for codon usage analysis. CODONFREQUENCY tabulates codon usage, CODONPREFERENCE predicts protein-encoding regions using the tabulated codon usage, and CORRESPOND utilises codon usage patterns to generate a matrix of Euclidean Manhattan distances between-genes. The formulae used by CORRESPOND (Equation 2-11), are related to those used during the preliminary stages of a correspondence analysis. However, the program name is somewhat misleading as CORRESPOND stops at this stage – equivalent to an early stage of a correspondence analysis (i.e. the generation of a matrix of distances), and does not actually perform a true analysis of correspondences.

Recently a world wide web based resource (URL http://acnuc.univ-lyon1.fr/mva/coa.html) was designed to facilitate the multivariate statistical analysis of codon usage (Perriere and Thioulouse 1996). This resource integrates WWW-Query (URL http://acnuc.univ-lyon1.fr/start.html) (Perriere and Gouy 1996), the web interface of the ACNUC database interrogation and sequence retrieval software (Gouy *et al*. 1985), with NetMul (URL http://biomserv.univ-lyon1.fr/base.html) a web interface for multivariate analysis methods (Thioulouse and Chevenet 1996). NetMul is a subset of the ADE (Analysis of Environmental Data) package (Thioulouse *et al*. 1995). These interfaces are integrated such that a group of sequences (i.e. a dataset) can be created using WWW-Query, the codon usage (or amino-acid usage) tabulated, and a MVA of the data performed using NetMul. The positions of the genes along the trends identified by the correspondence analysis are then displayed as a scatter plot (a GIF file which includes HTML map coordinates) (Thioulouse 1996).

There are limitations to using a web-based analysis system. As with all web-based resources, there is the problem of overall lack of bandwidth on the Internet. The interpretation of results from a MVA of codon or amino acid usage is difficult without reference to other properties that can often be used to quantify the trends. Sequence annotation is often incorrect and in many cases inconsistent, thus automatically generated datasets can include data of poor quality, which can severely distort the results of an analysis. All datasets must be carefully inspected. This web resource, in common with the other programs previously mentioned, does not attempt to identify potentially optimal codons.

Due to the limitations of the available analysis software and to facilitate my own analyses of codon usage variation, the program CodonW was envisaged. It was designed to simplify the analysis of codon usage, by integrating codon usage indices with multivariate statistical analysis (i.e. correspondence analysis) into a single program. It had to be simple to use and portable (in terms of both operating systems and machine architecture). With the increasing growth in the sequence databases, it was also designed not to have limits on the number of sequences that can be included in an analysis or sequence length.

Many codon-based analysis programs ignore the fact that there are alternatives to the universal genetic code. CodonW was designed to work with any variation of the genetic code. Seven alternatives to the universal genetic code have been built into CodonW; other codes can be added at compilation time, by editing a header file. Decisions regarding the translation of a codon, how synonymous each amino-acid is, the number of codons in a codon family, and which codons are synonyms are determined at run time.

Although the COA output from CodonW is more easily interpreted when presented graphically, CodonW does not have in-built graphics. There are numerous programs specifically designed to work with numerical data (e.g. Cricket Graph (MAC or PC), StatsView (MAC), Excel (MAC, PC), Minitab (MAC, PC, VMS, UNIX), SAS (PC, UNIX, VMS), SPSS (PC, UNIX, VMS), Harvard Graphics (PC), gnuplot (X-windows)) and therefore graphical presentation of the data is left to external software. All the programs mentioned above accept ASCII input files, by default CodonW creates much of its output in a tabulated format designed to be easily read by eye (output from a COA is always machine-readable). The command line switch (-machine) or the **Machine readable** option under the “Defaults menu” (Menu 3) changes the output format to a ASCII delimited format, which was designed to be machine-readable. The ASCII delimiter can be changed using the **Change ASCII delimiter in output** option under the “Defaults menu” (Menu 3).

2.2 System and Hardware

CodonW was developed under Digital-UNIX running on a DEC ALPHA AXP workstation. It has been ported to other flavours of UNIX, IBM compatible PCs (running Microsoft Windows

3.11 and Windows 95) and Apple Macintoshes (running System 7). CodonW can be used to calculate the base and dinucleotide composition of the complete *Haemophilus influenzae* genome (1.8-Mb), and the correspondence analysis of more than 2,000 sequences on both a PowerMac (System 7.5, 4-Mb RAM) and an IBM compatible PC (Microsoft Windows 95/98/NT, 8-Mb RAM).

CodonW is written in standard ANSI C (Kernighan and Richie 1988). It compiles cleanly using the GNU C compiler (version 2.7.2.1) with the stringent ANSI and pedantic command line switches. The CodonW source code, with instructions for installing can be downloaded as a compressed UNIX tar archive, by anonymous FTP from (URL ftp://molbiol.ox.ac.uk/cu/codonW.tar.Z). Additional documentation and a short tutorial on codon usage analysis can be found at the CodonW homepage (URL http://www.molbiol.ox.ac.uk/cu). The source code consists of seven C files and a header file. These files are included with this thesis as Appendix A. Alternatively, precompiled binaries are available at ftp://www.molbiol.ox.ac.uk/cu for Microsoft Windows 95 (compiled using Visual C++ version 4); Apple Macintosh (compiled using Code Warrior); Digital UNIX ; SunOS 5.6 and IRIX 6.2 (64 bit). The correspondence analysis routines used by CodonW are adapted from the NetMul library (Thioulouse and Chevenet 1996), a subset of ADE (Analysis of Environmental Data) package (Thioulouse *et al*. 1995).

2.3 Input files

Sequences to be analysed should be in a single file. They must be nucleic acid, protein encoding, sequential, and separated by at least one header line. A header line is defined as any line whose first character is either a semicolon ‘;’ or a right angled bracket ‘>’. There may be any number of header lines but they must precede each sequence, and the second or subsequent header lines are ignored. The first 20 characters after the ‘;’ or ‘>’ should be some unique description of the sequence as this is used as the sequence label in the output files. To avoid duplicate sequence labels each label is prepended with a sequential number. Those lines whose first character is neither ‘;’ or ‘>’, are considered to be sequence data. Sequences must be in the correct reading frame, and should not contain untranslated 5’ or 3’ sequence. CodonW assumes that the first character of the sequence is the first base of the first codon.

The format of each line of sequence data is relaxed, sequences can be either DNA or RNA, upper or lower case characters. Input lines may be any width and contain spaces and/or numbers. All non-alphabetical characters including white space (spaces, tabs, new-line) are stripped from the input. Each nucleotide must be represented by an IUB-IUPAC code. Nucleotides that cannot be identified must be represented by the IUPAC codes X or N. Codons that do not consist entirely of "A, T, U, C, or G" bases are considered non-translatable and are ignored.

CodonW performs some basic checks on the integrity of the input data. These include whether or not the sequence contains internal stop codons (usually an indicator that the sequence is out of frame), for the presence of non-IUPAC characters, and that the sequence is nucleic-acid. In *E. coli,* codon usage has been shown to vary within a gene (Eyre-Walker 1995a; Eyre-Walker and Bulmer 1993), variation in codon usage may be introduced by comparing partial and full-length sequences. Therefore, CodonW checks that each sequence contains a valid start and termination codon; any prokaryotic start codon (i.e. NTG, ATN) (Osawa *et al*. 1992) is considered valid. If any potential problems are found, warning messages are displayed, these can be redirected to a file. CodonW warns about potential problems with a sequence, but these sequences are still included in any analysis. For this reason, sequences that generate warnings should be carefully checked.

While CodonW normally requires only an input file, selection of some options will cause CodonW to prompt for additional files. The indices CAI, CBI and F_{op}, quantify the adaptation of codon usage towards a set of optimal codons. While optimal codons are known for some species and are in-built into CodonW, for most species they are not known. Therefore, selecting one of these indices will cause CodonW to prompt for a personal choice of optimal codons or codon relative adaptedness values (used by CAI). Before creating a file of relative adaptedness values, it is recommended that the original paper (Sharp and Li 1987a) be consulted. Input files which contain a personal choice of optimal codons are expected to contain a score for each codon: 1 for a rare codon; 3 for an optimal codon; 2 for other codons. These scores can be separated by any white space character, but must be in a pre-set sequence. That is, the first value must represent the codon UUU (1), the second UCU (2), with the last value representing GGG (64). For a complete list of these codes, see Table 2-1 for an explanation about how they are assigned see section 2.4 Data Recoding.

There is an alternative to creating these F_{op} or CAI input files by hand. During a correspondence analysis (of codon usage or relative synonymous codon usage - RSCU) the files “cbi.coa”, “fop.coa” and “cai.coa” are generated automatically. These files contain the codon usage information necessary for calculating their respective indices. However, to discover this codon usage information CodonW makes fundamental assumptions about the dataset, which must be verified by the researcher. These are:

- The dataset is a representative sample of genes that are lowly and highly expressed.
- The major trend in the codon usage is selection for optimal translation.
- The genes with the highest codon bias are the most highly expressed.
- All genes automatically assigned to the high bias subset are highly expressed.

The verification of these assumptions is not trivial. If the codon usage information in the files “cbi.coa”, “fop.coa” and “cai.coa” is accepted uncritically, it is possible that any indices calculated using them will be incorrect. When these assumptions are valid, these files greatly simplify the task of calculating these indices.

Another input file is prompted for, if **add additional genes after COA** is selected under the “advanced correspondence analysis” menu. Correspondence analysis can be adversely affected by genes that have atypical codon or amino-acid usage. Therefore, certain categories of genes (e.g. plasmid genes, transposons, etc.) are normally excluded from a COA. However, it is often useful to compare the codon usage of these excluded genes with those used in the COA. This is possible, by using a dataset of “good” genes (the main input file) to identify the trends in codon usage and then using these trends (or more precisely the vectors representing the trends) to transform the codon usage of the genes originally excluded, into axis co-ordinates. The two sets of axis coordinates are then directly comparable. For a more detailed, explanation see section 2.5.4 Correspondence analysis (Menu 5).

2.4 Data Recoding

CodonW internally recodes all nucleotides, codons, and amino-acids as numerical data. Nucleotides are recoded as T or U=1, C=2, A=3, G=4. Each codon is recoded as an integer value in the range 1 to 64, see Table 2-1. The 20 standard amino-acids and the termination codons are recoded as integer values in the range 1 to 21, see Table 2-2. The formula used to recode each codon is:

Equation 2-1

$$code = \left(\left(P_1 - 1\right) * 16\right) + P_2 + \left(\left(P_3 - 1\right) * 4\right) \quad 1 \leq code \leq 64$$

Each of three codon positions is represented by P1, P2, and P3. Thus, the codon ATG has the value 45.

$$code = \left(\left(3 - 1\right) * 16\right) + \; 1 + \left(\left(4 - 1\right) * 4\right) = 45$$

Non-standard nucleotides and non-translatable codons or the amino-acids that they represent are coded as zero.

2.5 The Menu Interface

CodonW was designed to be driven via a series of nested menus, each menu having its own online help. The initial menu allows the user to choose sub-menus to initiate an analysis or to exit the program. There are 12 menus in total with the deepest menu nested four deep, menus 2 and 6 are reserved for future use.

2.5.1 Loading sequence file (Menu 1)

Unless an input filename is given on the command line, the input file must be loaded using this menu. Three filenames are prompted for, the first of which is the main input file; it must exist, be

Table 2-1 Numerical values used for recoding codons

Code	Codon	**Amino-acid**	Code	Codon	**Amino-acid**	Code	Codon	**Amino-acid**	Code	Codon	**Amino-acid**
1	UUU	**Phe**	2	UCU	**Ser**	3	UAU	**Tyr**	4	UGU	**Cys**
5	UUC		6	UCC		7	UAC		8	UGC	
9	UUA	**Leu**	10	UCA		11	UAA	**STOP**	12	UGA	**STOP**
13	UUG		14	UCG		15	UAG		16	UGG	**Trp**
17	CUU		18	CCU	**Pro**	19	CAU	**His**	20	CGU	**Arg**
21	CUC		22	CCC		23	CAC		24	CGC	
25	CUA		26	CCA		27	CAA	**Gln**	28	CGA	
29	CUG		30	CCG		31	CAG		32	CGG	
33	AUU	**Ile**	34	ACU	**Thr**	35	AAU	**Asn**	36	AGU	**Ser**
37	AUC		38	ACC		39	AAC		40	AGC	
41	AUA		42	ACA		43	AAA	**Lys**	44	AGA	**Arg**
45	AUG	**Met**	46	ACG		47	AAG		48	AGG	
49	GUU	**Val**	50	GCU	**Ala**	51	GAU	**Asp**	52	GGU	**Gly**
53	GUC		54	GCC		55	GAC		56	GGC	
57	GUA		58	GCA		59	GAA	**Glu**	60	GGA	
61	GUG		62	GCG		63	GAG		64	GGG	

Table 2-2 Numerical values used to recode amino-acids.

Code	Amino-acid	One letter code	Code	Amino-acid	One letter code
1	Phe	F	2	Leu	L
3	Ile	I	4	Met	M
5	Val	V	6	Ser	S
7	Pro	P	8	Thr	T
9	Ala	A	10	Tyr	Y
11	Stop	*	12	His	H
13	Gln	Q	14	Asn	N
15	Lys	K	16	Asp	D
17	Glu	E	18	Cys	C
19	Trp	W	20	Arg	R
21	Gly	G			

readable, and contain sequence data. The format of the input sequences has been discussed above. The second file is used to store the results for indexes that can be easily summarised, that is those indices selectable using Menu 4 (see below). The suggested name for this file is the "root" of the input filename with the extension ".out". The third file is the bulk output file; this is used to record results or output which cannot be easily summarised into a single or few variables, that is output selectable using Menu 8. The suggested extension for the bulk output file is ".blk". If either of the two output filenames already exists, the user is prompted whether or not to overwrite, or quit. If they choose not to overwrite, they have the choice either to select a different filename or to append the output to the existing file. If the "silent" option has been selected (-silent on the command line or Menu 3 option 2) all files are overwritten without warning.

2.5.2 Changing the default values (Menu 3)

To improve flexibility, many of the default values used internally by CodonW (defined in the header file codonW.h) can be altered at runtime using this menu. Ten options can be customised.

Option (1) Change ASCII delimiter in output. The default ASCII delimiter used to separate information in machine-readable output files is a comma, which can be changed to either a tab or a space character.

Option (2) Run silently. This option can be used when running from a script file or as a batch job. If TRUE it suppresses warnings about overwriting files, the prompting for a personal choice of F_{op}, CBI or CAI values (although these can still be given via command line arguments), and the pause between pages of error or warning messages.

Option (3) Log warnings or information to a file. The default value for this option is FALSE, so all warning or error messages generated by CodonW are written to the screen via the standard error stream. If TRUE, the errors are redirected to a file, the filename of which is prompted for. This option is useful if there are a large number of sequences in the input file or there are many warning messages.

Option (4) Number of lines on screen. This is used to set the screen length, which is used during screen refreshing and pagination of error messages.

Option (5) Change the genetic code. By default, CodonW assumes the universal genetic code when translating and processing codons. This option allows alternative genetic codes to be selected. There are seven in-built alternatives. Genetic codes can be altered or added to by editing the structure codon_usage_struct in the header file codonW.h and modifying variable NumGeneticCodes. The available codes are:

(0)	Universal Genetic code	(TGA=*, TAA=*, TAG=*)
(1)	Vertebrate Mitochondrial code	(AGR=*, ATA=M, TGA=W)
(2)	Yeast Mitochondrial code	(CTN=*, ATA=M, TGA=W)
(3)	Filamentous fungi Mitochondrial code	(TGA=W)
(4)	Insects and *Plathyhelminthes* Mitochondrial code	(ATA=M, TGA=W, AGR=S)
(5)	Nuclear code of *Cilitia*	(UAA=Q, UAG=Q)
(6)	Nuclear code of *Euplotes*	(UGA=C)
(7)	Mitochondrial code of *Echinoderms*	(UGA=W, AGR=S, AAA=N)

Option (6) Change the F_{op} or CBI values. To calculate either the CBI or F_{op} indices, a set of optimal codons is required; the optimal codons of *E. coli* are assumed by default. This option, which displays a submenu, lists eight species where optimal codons have been identified. When calculating the F_{op} or CBI of genes from these species the appropriate set of codons should be selected. These codons are defined in the structure fop_struct in the header file codonW.h. Additional choices of optimal codon can be added to CodonW by editing this structure and modifying the variable NumFopSpecies. This is largely unnecessary as personal selections of optimal codons can be input at run-time. Optimal codons are available for the following species:

(0) *Escherichia coli*	(Ikemura 1985)
(1) *Bacillus subtilis*	(Sharp *et al*. 1990)
(2) *Dictyostelium discoideum*	(Sharp and Devine 1989)
(3) *Aspergillus nidulans*	(Lloyd and Sharp 1991)
(4) *Saccharomyces cerevisiae*	(Sharp and Cowe 1991)
(5) *Drosophila melanogaster*	(Shields *et al*. 1988)

(6) *Caenorhabditis elegans*	(Stenico, Lloyd and Sharp 1994)
(7) *Neurospora crassa*	

Option (7) Change the CAI values. In order to calculate the codon adaptation index it is necessary to assign a fitness value to each codon. The default fitness values are those of *E. coli*. Fitness values are very species-specific and using *E. coli* fitness values to calculate CAI values for other species is nonsensical. Before calculating the relative adaptedness (ω) values of a codon, a set of genes experimentally verified to be highly expressed must be identified. Such sets have been created for relatively few species. This menu lists the species where a reference set of highly expressed genes is known, and ω values assigned. These ω values are defined in the structure cai_struct in codonW.h. Additional selections of ω values can be added to CodonW by editing this structure and the variable NumCaiSpecies. Again, this is largely unnecessary as personal selections of fitness values can be input at runtime if calculating CAI. Values for codon ω values are available for:

(0) *Escherichia coli*	(Sharp and Li 1987a)
(1) *Bacillus subtilis*	(Shields and Sharp 1987)
(2) *Saccharomyces cerevisiae*	(Sharp and Cowe 1991)

Option (8) Toggle human or machine-readable output. The default format for CodonW output files is human-readable (COA analysis output files are always in machine-readable format). Machine-readable output is free-width data separated by an ASCII delimiter; this format is readily imported into a wide range of statistical and graphical analysis programs but is not easily read by eye. Human-readable output is more verbose but easier to read, and may contain additional information (see Figure 2-1). The output formats for codon usage, tabulation of codon usage, relative synonymous codon usage and base composition are those most radically affected by this option.

Option (9) Toggle output for each or all genes. By default, CodonW processes input data gene by gene. When the option "all genes" is selected, sequences are concatenated and processed as a single sequence. This option can be used to calculate the total codon or amino-acid usage; the

Machine readable output (truncated in width)

```
Gene_description,Len_aa,Len_sil,GC,GC3s,GCn3s,GC1,GC2,GC3,T1,T2,T3,C1,C2,C3,A1,A2,A3,G1
A00198.CYSE,273,261,0.556,0.594,0.538,0.623,0.432,0.612,0.125,0.289,0.260,0.223,0.253,0
.
```

Human readable output

```
Gene Name: G|A00198.CYSE 821 S. typhimurium cysE gene for serine acetyltransferase
Length   : 273 AA      Non-synonymous or synonymous codons ( 12 or   261)
 GC=0.556      GC3s=0.594     GC_not_GC3s=0.538
base    1       2       3       total          1       2       3       total
  T     0.125   0.289   0.260   0.225   W      0.377   0.568   0.388   0.444
  C     0.223   0.253   0.304   0.260   S      0.623   0.432   0.612   0.556
  A     0.253   0.278   0.128   0.220   R      0.652   0.458   0.436   0.515
  G     0.399   0.179   0.308   0.295   Y      0.348   0.542   0.564   0.485
```

Figure 2-1. An example of the machine and human-readable output formats of CodonW. A base composition analysis of *S. typhimurium cysE* is shown in both formats

average G+C content, F_{op}, etc. If concatenating DNA sequences and outputting in a fasta/Pearson format (tidy or reader output formats Menu 8), care should be taken to avoid frameshifts caused by partial codons. For other analyses, CodonW truncates any 3’ partial codons, to maintain the reading frame. As correspondence analysis of a concatenated sequence is nonsensical, COA is disabled when **all genes** is selected.

Option (10) Correspondence analysis defaults. This option allows access to the “advanced correspondence analysis” menu. This menu is normally accessed as a submenu from the “Correspondence analysis”(Menu 5), but is included here so that all runtime configurable options are accessible via this “Menu 3 Change default values” menu. The options selectable when using this sub-menu are dependent on whether COA will be of codon or amino-acid usage. If neither has been selected, the user is prompted to choose one. For a more information about this menu, see “Correspondence analysis (Menu 5)”.

2.5.3 Codon usage indices (Menu 4)

This menu is used to choose the indices calculated by CodonW as by default only the G+C content of the sequence is calculated. The calculation of these indices (except G+C content) is dependent on the genetic code selected under Menu 3. More than one index may be chosen at the same time.

Option (1) Codon Adaptation Index (CAI) (Sharp and Li 1987a). CAI is a measurement of the relative adaptedness of the codon usage of a gene towards the codon usage of highly expressed genes. The relative adaptedness (ω) of each codon is the ratio of the usage of each codon, to that of the most abundant codon within the same synonymous family. The relative adaptedness of codons, for albeit a limited choice of species, can be selected from Menu 3. The user can also input a personal choice of values (see above). The CAI index is defined as the geometric mean of these values. Non-synonymous codons and termination codons (genetic code dependent) are excluded. To aid computation, the CAI is calculated as

Equation 2-2

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

Where ω_k is the relative adaptedness of the k^{th} codon and L is the number of synonymous codons in the gene.

To prevent a codon which is absent from the reference set, but present in other genes, from having a relative adaptedness value of zero, (which could result in a CAI of zero); it was suggested that absent codons should be assigned a frequency of 0.5 when estimating ω (Sharp and Li 1987a). An alternative suggestion was that ω should be adjusted to 0.01 where otherwise it would be less than this value (Bulmer 1988). CodonW does not adjust the ω value unless ω is zero or effectively zero (<0.0001) in which case values are assigned as being 0.01.

Option (2) Frequency of Optimal Codons (F_{op}) (Ikemura 1981a). F_{op} is the fraction of synonymous codons (genetic code-dependent) which are optimal codons. Optimal codons for several species are in-built and can be selected using Menu 3. By default, the optimal codons of *E. coli* are assumed. The user may also enter a personal choice of optimal codons. If rare synonymous codons have been identified, there is a choice of calculating the original F_{op} index (Equation 2-3) or a modified F_{op} index (Equation 2-4). Fop values for the original index are always between 0 (where no optimal codons are used) and 1 (where only optimal codons are used). When calculating a modified F_{op} index, negative values are possible but these are adjusted to zero.

Equation 2-3

$$Fop = \frac{\mathrm{N}_{optimal\_codons}}{\mathrm{N}_{synonymous\_codons}}$$

Equation 2-4

$$Fop_{(\mathrm{mod})} = \frac{\left(\mathrm{N}_{optimal\_codons} - \mathrm{N}_{rare\_codons}\right)}{\mathrm{N}_{synonymous\_codons}}$$

Where N is the frequency each codon type is used.

Option (3) Codon Bias Index (CBI) (Bennetzen and Hall 1982). Codon bias index is a measure of directional codon bias, and is a measure of the extent to which a gene uses a subset of optimal codons (see Equation 2-5).

Equation 2-5

$$CBI = \frac{N_{opt} - N_{ran}}{N_{tot} - N_{ran}}$$

Where N_{opt} = number of optimal codons; N_{tot} = number of synonymous codons; N_{ran} = expected number of optimal codons if codons were assigned randomly. CBI is similar to F_{op} as used by Ikemura, with N_{ran} used as a scaling factor. In a gene with extreme codon bias, CBI will equal 1.0, in a gene with random codon usage CBI will equal 0.0. Note that it is possible for N_{opt} to be less than N_{ran}, which results in a negative value for CBI.

Option (4) The effective number of codons (EN_C) (Wright 1990). This index is a simple measure of overall codon bias and is analogous to the effective number of alleles measure used in population genetics. Knowledge of the optimal codons or a reference set of highly expressed genes is unnecessary when calculating this index.

Initially the homozygosity for each amino acid is estimated from the squared codon frequencies (see
Equation 2-6).

Equation 2-6

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

Where k= number of synonyms; *n*= total usage of k-fold synonymous amino-acid; $\hat{F}$ = homozygosity; p_{i} = frequency of usage of synonymous codon *i*.

The universal genetic code has five amino-acid family types (non-synonymous, 2-fold, 3-fold, 4-fold and 6-fold synonymous amino-acids). The arithmetic average of all non zero homozygosity values for each family type is calculated and the contributions from each of the synonymous families is summed to calculate Nc (see Equation 2-7).

Equation 2-7

$$\hat{N}_c = 2 + \frac{9}{\bar{\hat{F}}_2} + \frac{1}{\hat{F}_3} + \frac{5}{\bar{\hat{F}}_4} + \frac{3}{\bar{\hat{F}}_6}$$

CodonW uses a modified version of this equation that is genetic code independent (see Equation 2-8).

Equation 2-8

$$\hat{N}_c = N_1 + \sum_{i=2}^{x} \frac{N_i}{\bar{\hat{F}}_i}$$

Where x= number of members in the most synonymous family; N_1= frequency of non-synonymous amino-acids; N_i= number of *i*-fold amino-acid families with homozygosity greater than 0; N_C is summed over all synonymous amino-acid families.

If amino-acids are rare or missing, adjustments must be made. For absent amino-acids the numerator is decreased, to reciprocally average for only the amino-acids present. For sequences where a synonymous family of amino-acids is empty ($\hat{F}_n$=0), N_C is not calculated as the gene is assumed to be either too short or to have extremely skewed amino-acid usage (Wright 1990). An exception is made for genetic codes where isoleucine is a 3-fold synonymous amino-acid and is absent from the protein. Then $\hat{F}_3$ is estimated as the average of $\bar{\hat{F}}_2$ and $\bar{\hat{F}}_4$. The reported value for EN_c is always between 20 (when effectively only a single codon is used for each amino-acid) and 61 (when codons are used randomly). If EN_c is greater than 61 (because codon usage is more evenly distributed than expected), it is corrected to 61.

Option (5) G+C content of the gene. This is calculated to be the fraction of nucleotides that are guanine or cytosine.

Option (6) G+C content at 3rd position of synonymous codons ($GC3_S$). This option calculates the fraction of codons synonymous at the third codon position, and having guanine or cytosine at that third position.

Option (7) Silent base composition. This option calculates four indices, $G3_S$, $C3_S$, $A3_S$ and $T3_S$. Although correlated with $GC3_S$, this index is not directly comparable. It quantifies the usage of each nucleotide at synonymous third codon positions as a proportion of the maximum usage of that nucleotide could have without altering the amino acid composition. A similar index of this type has been used to demonstrate a correlation in *Caenorhabditis elegans* between codon usage variation and synonymous codons with G in the third position. Though it was not clear whether it reflected variation in base composition, mutational biases in different regions of the *C. elegans* genome, or another factor (Stenico, Lloyd and Sharp 1994).

Options (8+9) Number of silent sites (L_{Sil}) and amino acids (L_{AA}) are self-explanatory.

Option (10) Hydrophobicity of protein (GRAVY). This calculates the general average hydrophobicity (GRAVY) score for the conceptually translated gene product. It is calculated as the arithmetic mean of the sum of the hydrophobic indices of each amino-acid (Equation 2-9) (Kyte and Doolittle 1982). This index has been previously used to quantify the major COA trends in the amino-acid usage of *E. coli* genes (Lobry and Gautier 1994).

Equation 2-9

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

Where N is the number of amino-acids, and K_i is the hydrophobic index of the i^{th} amino-acid.

Option (11) Aromaticity score. This is the frequency of aromatic amino-acids (Phe, Tyr, Trp) in the hypothetical translated gene product, and as given by Equation 2-10.

Equation 2-10

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

Where v_i is either 1 (for an aromatic amino-acid) or zero, N is the total of amino-acids. This index has also been previously used to quantify the major COA trends in the amino-acid usage of *E. coli* genes (Lobry and Gautier 1994).

2.5.4 Correspondence analysis (Menu 5)

Correspondence analysis (COA) is a sensitive method for the exploration of the non-random usage of synonymous codons found in many organisms. It has identified trends in the codon usage of *E. coli* (Holm 1986; Medigue *et al*. 1991), *B. subtilis* (Sharp *et al.* 1990; Shields and Sharp 1987), *S. cerevisiae* (Sharp and Cowe 1991), *D. melanogaster* (Shields *et al*. 1988), *Dictyostelium discoideum* (Sharp and Devine 1989), *C. elegans* (Stenico, Lloyd and Sharp 1994), *Rickettsia prowazekii* (Andersson and Sharp 1996a), *B. burgdorferi* (McInerney 1998) and *M. tuberculosis* (Andersson and Sharp 1996b).

Correspondence analysis requires contingency tables, which are counts of the joint occurrences of rows and columns of a table. Therefore, sequence data must be transformed, the frequency of each codon (or amino acid) is tabulated, and this is then converted into an Euclidean measurement of distance between the rows or columns. A common distance measurement used in COA is as shown in Equation 2-11.

Equation 2-11

$$d^2(i_1 i_2) = \sum_{j=1}^{X} \left(f_{i_1 j} - f_{i_2 j} \right)^2$$

Where, for COA of codon usage, $f_{i_1 j}$ and $fi_2 j$, are the frequencies of codon j in sequences i_1 and i_2. X is the number of codons over which the distance is to be totalled (when only synonymous codons are analysed X=59).

The GCG program, CORRESPOND calculates a similar distance measurement, summed over all 64 codons, but limits its analysis to the calculation of this distance matrix.

The above distance measurement is unscaled; rare codons have little influence on the distance and are effectively ignored. To correct for this, CodonW uses a scaled distance measurement (Equation 2-12) as recommended by Grantham and co-workers (Grantham *et al*. 1981).

Equation 2-12

$$d^2(i_1 i_2) = \sum_{j=1}^{X} \frac{1}{n_j} \left(f_{i_1 j} - f_{i_2 j} \right)^2$$

Where n_j is the number of occurrences of codon j in the sample; f_{i_1j} and $fi2j$ are the frequencies of codon j in sequences $i1$ and $i2$; X is the number of codons over which the distance is measured.

Analysis of a large number of distances would ordinarily be very time consuming. Correspondence analysis provides a simple visualisation of these distances by projecting the points from their original multidimensional space onto lower dimensions, with genes with similar distances plotted as neighbours (Grantham *et al*. 1981). In addition to calculating the coordinates for the projection of these points, correspondence analysis (as implemented in CodonW) also reports the total inertia of the data, the eigenvalue and the relative inertia explained by each axis, the absolute and relative contribution of each gene, codon or amino-acid on each axis.

The correspondence analysis menu (Menu 5) has four options, the default option being not to generate a correspondence analysis (Option 4).

Option (1) Correspondence analysis of codon usage. This generates a correspondence analysis of codon usage. By default, only synonymous codons are included, however the codons that are included or excluded can be changed using the advanced COA menu (see below). When analysing synonymous codon usage, the analysis has 58 degrees of freedom (or N-1 if the number of genes (N) is less than 59).

Option (2) Correspondence analysis of RSCU. This generates a correspondence analysis of relative synonymous codon usage (RSCU). By default, only synonymous codons are analysed. RSCU is calculated as the ratio of the observed frequency of a codon to the frequency expected under unbiased usage within a synonymous codon group (see below) (Sharp, Tuohy and Mosurski 1986). Correspondence analysis of RSCU removes variation caused by unequal usage of amino-acids but the degrees of freedom are reduced to 40 (or N-1 if the number of genes (N) is fewer than 41).

Option (3) Correspondence Analysis of Amino-Acid Usage. This generates a correspondence analysis of amino-acid composition. By default, the 20 standard amino-acids are analysed.

Amino-acids can be included or excluded using the advanced COA menu. There are 19 or N-1 degrees of freedom.

Option (4) Do not perform a correspondence analysis. This is the default option.

2.5.5 Advanced Correspondence analysis

When a correspondence analysis option is chosen, there is a further choice of whether or not to access the advanced COA menu. This advanced menu allows much greater control during correspondence analysis.

Option (1) Unselect or select. Depending on whether correspondence analysis is of amino-acid or codon usage, this menu allows the selection of which codons or amino-acids are to be analysed. This allows the user to override the default selections. If the COA is of codons, non-synonymous codons and termination codons are by default excluded. Excluded codons or amino-acids are indicated by square brackets, see Figure 2-2. Codons or amino acids can be selected or unselected by entering their index number.

Option (2) Change the number of axes. The number of axes (or trends) generated by a correspondence analysis is N-1, where N is either the number of genes or columns (whichever is the lesser). However, as the most important components of data inertia are explained by the first axis and progressively less by each subsequent axes the default is to record the ordination for the first four. This option allows the user to record coordinates on any number of axes up to the total number of axes generated.

Option (3) Add additional genes after correspondence analysis. Each axis generated by correspondence analysis is represented by a multidimensional vector. The ordination of a gene on each axis is the product of the gene's codon usage and the axis vector. As the vector is itself a product of the codon usage, the vectors can be affected by unusual codon usage. For example, an analysis of both

```
 ( 1) Phe       UUU  ( 2) Ser   UCU  ( 3) Tyr   UAU  ( 4) Cys   UGU
 ( 5)           UUC  ( 6)       UCC  ( 7)       UAC  ( 8)       UGC
 ( 9) Leu       UUA  (10)       UCA [(11) TER   UAA][(12) TER   UGA]
 (13)           UUG  (14)       UCG [(15)       UAG][(16) Trp   UGG]

 (17)           CUU  (18) Pro   CCU  (19) His   CAU  (20) Arg   CGU
 (21)           CUC  (22)       CCC  (23)       CAC  (24)       CGC
 (25)           CUA  (26)       CCA  (27) Gln   CAA  (28)       CGA
 (29)           CUG  (30)       CCG  (31)       CAG  (32)       CGG

 (33) Ile       AUU  (34) Thr   ACU  (35) Asn   AAU  (36) Ser   AGU
 (37)           AUC  (38)       ACC  (39)       AAC  (40)       AGC
 (41)           AUA  (42)       ACA  (43) Lys   AAA  (44) Arg   AGA
[(45) Met       AUG] (46)       ACG  (47)       AAG  (48)       AGG

 (49) Val       GUU  (50) Ala   GCU  (51) Asp   GAU  (52) Gly   GGU
 (53)           GUC  (54)       GCC  (55)       GAC  (56)       GGC
 (57)           GUA  (58)       GCA  (59) Glu   GAA  (60)       GGA
 (61)           GUG  (62)       GCG  (63)       GAG  (64)       GGG
```

Figure 2-2. Simple menu interface used to select or unselect the codons that are to be included in the correspondence analysis. This menu can be selected using the advanced COA menu. The codons surrounded with square brackets are currently excluded.

chromosomal and plasmid genes is often difficult, the codon usage of one will perturb the other. Each dataset can be analysed separately but because axes vectors would not be the same, it would be difficult to make comparisons between the analyses. To circumvent this, it is necessary to generate the COA vectors using one dataset and apply these vectors to another, thus direct comparison between the ordination of genes is possible using this option.

The user is prompted for the file containing the additional sequences to which the vectors are to be applied. The vectors are calculated as normal, using the genes contained in the standard input file (Menu 1). The co-ordinates and any additional information about these original genes are recorded, then the additional genes are read-in, and the original vectors applied to them. The ordinations of these additional genes are then appended to the COA output files (for more information about COA output files see below).

Option (4) Toggle level of correspondence analysis output. By default this option is set to "normal" but can be toggled to "exhaustive". If the exhaustive output option is selected, then in addition to the normal information about gene and codon or amino-acid ordination, additional information about inertia of the rows and columns is generated. This additional information includes the absolute contribution that the inertia of each row (gene) or column (codon or amino-acid) has on the recorded axes. It also includes how much of the variation of each row or column is explained by each of the recorded trends.

Option (5) Change number of genes used to identify optimal codons. Correspondence analysis of codon usage or RSCU, where the major trend in codon usage is selection for optimum translation, can be used to identify optimal codons. This is achieved by contrasting the codon usage of two groups of genes, composed of the genes that lie at either end of the principal trend (axis 1). By default, these are the top and bottom 5% of genes (based on axis 1 ordination). This option allows the size of these groups to be changed, it can either be set as a percentage or an absolute number of genes.

2.5.6 Output options in CodonW (Menu 8)

CodonW results that are not related to correspondence analysis, and which cannot be easily summarised as a few variables, are written to the “bulk” output file. This menu is used to select what results are to be written to this output file. There are 10 options and multiple options cannot be simultaneously selected. To facilitate those occasions when more than one option is required, each time this menu is selected there is a prompt for an alternative output filename or the output can be appended to the current output file.

Option (1) Fasta format output of DNA sequence. Sequences are reformatted into a fasta or Pearson format, with a single header line per sequence. The header has a fixed width, i.e. the angle bracket, sequence name and gene lengths are written at fixed positions. Sequence data is output with 60 characters per line.

Option (2) Reader format output of DNA sequence. This format is similar to the standard fasta format, except nucleotides are written in groups of three separated by a space (i.e. as codons). This is fixed width format to be backwardly compatible with legacy FORTRAN code.

Option (3) Translate input file to amino-acids. This option conceptually translates nucleotide sequence data into amino-acids, using the current genetic code (specified under Menu 3). Amino-acid sequence is output in a Fasta or Pearson format.

Option (4) Codon Usage. This is the default bulk output option. The frequency of each codon is output as four rows with 16 columns per row. The codons are written in numerical order, left to right, according to their occurrence in Table 2-1. If the output format is “Human-readable” this option creates tables of codon usage including descriptions and RSCU values for each codon see Figure 2-3

Option (5) Amino-acid usage. This option generates tables showing frequency of each amino-acid, untranslatable codons and stop codons, 1 row per gene and 23 columns per row. The first column contains a unique label for the gene, the second records the untranslatable codons, the third and subsequent columns summarize the amino-acid and termination codon usage. The

```
Phe UUU   13 0.70 Ser UCU    4 0.50 Tyr UAU    7 0.48 Cys UGU    1 0.50
    UUC   24 1.30     UCC    5 0.63     UAC   22 1.52     UGC    3 1.50
Leu UUA    1 0.11     UCA    7 0.88 TER UAA    0 0.00 TER UGA    1 3.00
    UUG   10 1.05     UCG    9 1.13     UAG    0 0.00 Trp UGG    3 1.00

    CUU    4 0.42 Pro CCU    8 0.97 His CAU    1 0.40 Arg CGU   10 0.98
    CUC   13 1.37     CCC   13 1.58     CAC    4 1.60     CGC   36 3.54
    CUA    5 0.53     CCA    4 0.48 Gln CAA   21 1.05     CGA    3 0.30
    CUG   24 2.53     CCG    8 0.97     CAG   19 0.95     CGG   12 1.18

Ile AUU   15 1.18 Thr ACU    5 0.53 Asn AAU    7 0.58 Ser AGU   10 1.25
    AUC   22 1.74     ACC   20 2.11     AAC   17 1.42     AGC   13 1.63
    AUA    1 0.08     ACA    1 0.11 Lys AAA   18 1.13 Arg AGA    0 0.00
Met AUG    4 1.00     ACG   12 1.26     AAG   14 0.88     AGG    0 0.00

Val GUU    9 0.78 Ala GCU   13 0.84 Asp GAU   11 0.88 Gly GGU   18 1.41
    GUC   18 1.57     GCC   26 1.68     GAC   14 1.12     GGC   28 2.20
    GUA    1 0.09     GCA   11 0.71 Glu GAA   27 1.17     GGA    1 0.08
    GUG   18 1.57     GCG   12 0.77     GAG   19 0.83     GGG    4 0.31

684 codons in ANAPCE.APCE (used Universal Genetic code)
```

Figure 2-3. Example of the human-readable RSCU and "Count codon usage" output from CodonW.

output is written in the same sequential order as defined in Table 2-2. Human-readable output format is tabulated frequency data, which included the IUB-IUPAC three letter amino-acid code.

Option (6) Relative Synonymous Codon Usage (RSCU). Relative synonymous codon usage is calculated as the ratio of observed codon frequency to the frequency expected if codon usage was uniform (see Equation 2-13). RSCU output has 4 rows per gene with 16 columns per row, with the output order left to right in the same order as defined in Table 2-1. When the output format is set to "Human-readable" the output is similar to the "Human-readable" output from Option 4 see Figure 2-3.

Equation 2-13

$$RSCU_x = \frac{\text{Frequency of codon}_\text{x}}{\text{Expected frequency of codon}_\text{x}\text{ if codon usage was uniform}}$$

Option (7) Relative Amino-Acid Usage (RAAU). This calculates the frequency of each amino-acid relative to the total number of amino-acids in each gene. The format of the output is the similar to Option 5 "Amino-Acid usage".

Equation 2-14

$$RAAU_x = \frac{\text{Frequency of amino acid}_\text{x}}{\text{Total number of amino acids}}$$

Option (8) Dinucleotide frequencies. The frequency of dinucleotides is recorded for the overall sequence and for each of the three reading frames. The human and machine-readable output formats are similar, "Human-readable" output is recorded with one row per frame, with 18 columns per row (sequence label, the frame and 16 values). "Machine-readable" records all the output as a single row.

Option (9) Base composition analysis. This option records nucleotide frequency at each codon position. It also reports GC, GC_{3s}, and GCns (G+C content excluding nucleotides used to calculate GC_{3s}). If the output format is "Human-readable", W (A+T), S (G+C), R (A+G) and Y

(T+C) are also calculated for each codon position, the output is also more verbose (see Figure 2-1).

Option (10) No output written to file. This is useful when working with large datasets and when disk storage is limited. This option suppresses all output to the “bulk” file.

2.6 Other Menus

There are two additional menus accessible from the main menu; these are Menu 7 “Learn the code” and Menu 9 “About CodonW”; neither is directly relevant to a codon usage analysis. Menu 7 was designed to help learn the various genetic codes, and it asks questions about codon usage and translation. The genetic code taught could be any one selectable using Menu 3 “Change defaults”. Menu 9 prints a welcome banner, author name, version number, and the date of compilation.

2.7 Command Line options

The menu interface was designed to simplify an analysis of codon usage but frequent users, those running in batch mode, or those trying to integrate with a graphical user interface (GUI) will find the CodonW command-line interface very useful. There are 43 command-line options and these are listed with descriptions in Table 2-3. All CodonW menu options are available. There are

three options unique to the command-line, these are -help (print a summary of all command-line options), -nomenu (by-pass the menu interface) and -nowarn (switch off all warning messages). There are some features available using the menu interface but not the command-line. These are Menu 7 and Menu 9 (see above), the redirection of warning messages to file, the selection or exclusion of codons or amino-acids from a COA and the “Add additional genes after correspondence analysis” advanced COA options.

A command line option is considered as any string whose first character is a dash, after the valid options (those listed in Table 2-3) have been processed, any remaining options are reported as errors. Any remaining command-line parameters are considered to be the input

Command-line only options

Command line switch	Description
-h(elp)	Print summary of command line options
-nomenu	Prevent the menu interface being displayed
-nowarn	Suppress display of sequences warnings

Defaults

-silent	Overwrite existing files silently, don’t prompt for output filenames
-totals	Concatenate genes into a single sequence (useful for calculating average values)
-machine	Select Machine readable output
-human	Select Human readable output
-code N	Genetic code as defined under Menu 3 Option 5
-f_type N	F_{op} or CBI codons as defined by Menu 3 Option 6
-c_type N	CAI fitness values as defined by Menu 3 Option 7
-t[char]	Column separator used in machine readable files

Codon and amino-acid usage indices

-cai	Calculate Codon Adaptation Index (CAI)
-fop	Frequency of OPtimal codons index (FOP)
-cbi	Codon Bias Index (CBI)
-enc	Effective Number of Codons (EN_c)
-gc	G+C content of gene (all codon positions)
-gcs3	G+C of synonymous codons 3rd positions
-sil_base	Base composition at synonymous third codon positions
-N_sil	Number of synonymous codons
-N_aa	Number of synonymous and non-synonymous codons
-aro	Predict protein aromaticity
-hyd	Predict protein hydrophobicity
-all_indices	Calculate all the above indices
-cai_file [filename]	User inputted file of CAI values
-cbi_file [filename]	User inputted file of CBI values
-fop_file [filename]	User inputted file of F_{op} values

Correspondence analysis (COA) options

-coa_cu	COA of codon usage frequencies
-coa_rscu	COA of Relative Synonymous Codon Usage
-coa_aa	COA of amino-acid usage frequencies
-coa_expert	Generate detailed (expert user) statistics on the correspondence analysis
-coa_axes N	Number of axis or trends to record
-coa_num N	Number of genes used to identify optimal codons. N may be a percentage.

Bulk output options (only one can be selected per analysis)

-aau	Amino-acid Usage (AAU)
-raau	Relative Amino-acid Usage
-cu	Codon Usage (CU) (default)
-cutab	Tabulation of codon usage
-rscu	Relative Synonymous Codon Usage (RSCU)
-fasta or –tidy	Reformat input file into fasta format
-reader	Reader format (codons are separated by spaces)
-transl	Conceptual translation of DNA into amino-acid
-base	Detailed report of codon G+C composition
-dinuc	Dinucleotide usage
-noblk	No output to be written to the bulk output file

Table 2-3. CodonW command-line switches with a brief explanation, arranged according to the menus they replace. N is an integer, char is a character, and filename is the name of an input or output file.

-silent and -nomenu options are used together, the default filenames are used for the output filenames.

2.7.1 By another name

CodonW can emulate a range of useful utility codon usage and sequence analysis programs. These programs perform the same tasks as CodonW, but several menu items or command-line options may need to be chosen. In essence, CodonW checks the command-line to ascertain what name was used to invoke it. If this name matches a list of accepted program names, it will emulate that particular program or more precisely it assumes that certain command-line options have been given (for a list of names with descriptions of the emulated program see Table 2-4). The purpose of this feature is to further simplify program usage by eliminating both command-line and menu interfaces. However, the command-line is not completely redundant, options such as concatenating genes (-totals), changing the choice of optimal codons (-f_type), selecting between machine (-machine) or human (–human) readable output or selecting an alternative genetic code (-code) are still valid.

2.8 Correspondence analysis in CodonW.

The most complex analyses that CodonW performs, are correspondence analysis of either amino-acid or codon usage. In essence, correspondence analysis creates a series of orthogonal axes to identify trends to explain the data variation, with each subsequent axis explaining a decreasing amount of the variation (Benzecri 1992). Correspondence analysis then assigns an ordination for each gene and codon (or amino-acid) onto each axis. A correspondence analysis generates a large quantity of data. CodonW writes the core data necessary to interpret the COA to the file “summary.coa”. To aid analysis most of this information is also duplicated as separate output files with the extension “.coa”, for a description of their contents see Table 2-5.

Optimal codons have been defined as those codons which occur more often (relative to their synonyms) in highly expressed genes, compared with lowly expressed genes (Ikemura 1981a). CodonW uses a modification of this definition, where optimal codons are defined as those

Program name	Description
cu	Count codon usage
cutab	Count and tabulate codon usage
rscu	Calculate relative synonymous codon usage (RSCU)
cutot	Count and tabulate overall codon usage, amino-acid and RSCU usage
aau	Count amino-acid usage
raau	Calculate relative amino-acid usage
tidy	Reformat sequences into Fasta or Pearson format
reader	Reformat input sequence data into space codons
transl	Conceptually translate sequence in to amino-acids
bases	Calculates base composition in each reading frame
gc3s	Calculates G+C content of synonymous third codon positions
gc	Calculates G+C content of each gene
base3s	Calculate base composition of synonymous third codon positions
dinuc	Count dinucleotide frequency
cai	Calculates codon adaptation index
fop	Calculates frequency of optimal codons
cbi	Calculates codon bias index
enc	Calculates effective number of codons

Table 2-4. If CodonW is renamed or called in such a way that the first argument is one of a pre-determined selection, CodonW will emulate that particular program. The accepted program names and a brief description of each is given.

Filename	Description of contents
summary.coa	This file contains a summary of all the information generated by correspondence analysis, including all the data written to files listed below, except for the output written to cusort.coa.
eigen.coa	Each axis generated by the correspondence analysis is represented by a row of information. Each row consists of four columns, the axis number, the axis eigenvalue, the relative inertia of the axis, and the sum of the relative inertia.
amino.coa† or codon.coa	Each codon or amino-acid included in the correspondence analysis is represented by a row. The first column is description of the variable, the subsequent columns contain the coordinate of the codon or amino-acid on each axis, the number of axes is user definable.
genes.coa	Each row represents one gene, the first column contains a unique description for each gene, subsequent columns contain the genes coordinate on each of the recorded axes. If additional genes are added to the correspondence analysis (advanced correspondence analysis option), the coordinates of these genes are appended to this file.
cusort.coa†	Contains the codon usage of each gene, sorted by the gene's coordinate on the principal axis, this information is used to generate the table recorded to the file hilo.coa.
hilo.coa†	This files records the results of 2 way Chi-squared contingency test between the codon usage of two groups of genes (the composition of which is determined by the "advanced correspondence analysis option") sampled from the extremities of axis 1 as recorded in cusort.coa.
cai.coa†	Contains the relative usage of each codon within each synonym family, the most frequent codon is assigned the value one and the usage of other codons within the same synonym family are expressed relative to this. This file can be used to calculate CAI with a personal choice of values.
fop.coa†	Contains a list of the optimal codons and non-optimal codons as identified in the file "hilo.coa". The format of this file can be utilised by CodonW to calculate F_{op} using a personal choice of optimal codons.
inertia.coa	This file is generated when the exhaustive output option is selected under the advanced correspondence analysis menu. It reports the absolute contribution of each gene and codon or amino-acid, to the inertia explained by each of the axes. It also records the portion of inertia of each gene and codon or amino-acid explained by each trend.

Table 2-5 Description of output files created by a CodonW correspondence analysis. † Files that are not generated during the correspondence analysis of amino-acids.

levels (Lloyd and Sharp 1991; Lloyd and Sharp 1993; Sharp and Cowe 1991; Sharp *et al*. 1988; Shields and Sharp 1987). Significance is assessed by a two-way chi square contingency test with the criterion of p <0.01. The advantage of this test is that differences in codon usage between highly and lowly expressed genes caused by random noise are suppressed. A disadvantage is that significance is dependent on sample size.

If the major trend in the variation of codon or RSCU usage is correlated with gene expression it is possible to identify the codons used preferentially in highly expressed genes, i.e. the optimal codons. To simplify analysis CodonW assumes that the principal trend is correlated with gene expression, it then uses this assumption to identify optimal codons. Although CodonW automatically generates putative optimal codons, these should not be accepted until it has been established that the principle trend in codon usage variation is selection for optimum translation. It is the researcher's responsibility to establish this before accepting the putatively identified codons as truly optimal. This raises the problem of how to establish whether or not the major trend is correlated with expression. While not definitive, there are some useful indicators and these are:

(1) If selection for optimum translation is the major trend in codon usage variation, then genes found at one extreme of the principal axis (axis orientation is arbitrary), should include the ribosomal proteins and glycolytic enzymes and those at the opposite end would be expected to be lowly expressed (e.g. regulatory proteins and cytoplasmic membrane proteins).

(2) The principal axis should explain a large proportion (>15%) of the total variation in the data and approximately two times as much of the variation as the second and subsequent axis.

(3) A subset of codons (UUC, UAC, AUC, AAC, GAC and GGU) are optimal in *E. coli, B. subtilis, S. cerevisiae, S. pombe,* and *D. melanogaster* (Sharp and Devine 1989), and in species that have selection for optimal translation they are usually optimal codons. Similarly, certain codons are often avoided in highly expressed genes (in prokaryotes AGG and AGA are usually rare). The putative optimal codons would be expected to agree with these observations.

(4) The ordination of the genes on the principal axis should be significantly correlated with some independent measure of codon bias such as the effective number of codons EN_C.

(5) The possibility that the principle trend is due to a mutational bias such as the variation in the G+C content (the principal trend in the codon usage of many eukaryotes), must always be investigated and eliminated. This is particularly important where there is a significant correlation between some measure of base composition (such as $GC3_S$ or GC) and the principal axis.

CodonW assumes that the trend represented by axis 1 is correlated with expression, and attempts to identify optimal codons, but the adage GIGO "garbage in, garbage out" must be stressed. If the trend represented by axis 1 is not correlated with expression then the identified codons are not optimal codons.

To identify the putative optimal codons, genes are sorted according to their position on the principal axis and their sorted codon usage is recorded in the file "cusort.coa". Two groups are read, taken from the start and end of this file (i.e. equivalent to the ends of the principal axis). By default, the number of genes in each group is 5% of the total number of genes in the dataset but this can be altered by using the advanced correspondence analysis menu or with the command-line option "–coa_num". The codon usage of each group of genes is summed.

A prerequisite for the identification of optimal codons is the identification of the highly expressed genes. As the orientation of the principal axis is arbitrary, the highly expressed genes could be either group of genes. Fortuitously, in species where there is codon selection for optimum translation, this selection increases codon bias in the more highly expressed genes. It is possible to exploit this by measuring the codon bias of both sets of genes using EN_C. The group of genes with the lower EN_C (the genes with the strongest codon bias) are putatively considered to be expressed at a higher level than the group with a higher EN_C (this does not always hold for species with extreme G+C mutational biases e.g. *Dictyostelium discoideum* where the set of genes with the lower ENc have optimal codon usage).

Once CodonW has completed its two way chi-squared test on the codon pairs from these two groups of genes, their codon usage and RSCU are output as a table in “summary.coa” and “hilo.coa” respectively. Codons putatively identified as optimal ($p<0.01$) are indicated with an asterisk (*). Although not automatically considered optimal by CodonW, codons that occur more frequently in the highly expressed dataset at $0.01<p<0.05$ are indicated with @. Codons that have an RSCU frequency less than 0.10 are arbitrarily considered as rare and are indicated with a minus ′-′. The Chi values for all significant codons are also recorded.

CodonW calculates three indices that measure the degree to which gene codon usage has adapted towards a set of optimal codons. These are the frequency of optimal codons (F_{op}), the codon bias index (CBI) and the codon adaptation index (CAI). To calculate these indexes, information about codon usage in the species being analysed is required. The indices F_{op} and CBI used the optimal codons for the species, while the index CAI uses codon adaptation values. For those species where this information is known, it has been in-built into CodonW. For other species, the indices should not be calculated unless the user supplies additional information via supplementary input files.

Once optimal codons and a group of highly expressed genes have been putatively identified, CodonW uses this information to calculate relative adaptedness values, which are recorded in “summary.coa” and “cai.coa”. Relative adaptedness values are calculated as the ratio of the usage of each codon relative to the most abundant codon for each synonymous family, in a set of highly expressed genes. Missing codons are assigned the frequency 0.5. The optimal codons are also recorded in “cbi.coa” and “fop.coa”. Again, it must be stressed that CodonW makes a number of critical assumptions to create these files as previously listed in section 2.3. If these assumptions are valid then the files “cai.coa”, “cbi.coa” and “fop.coa” can be used when prompted for a “personal choice of values”, to calculate the indexes CAI, CBI and F_{op} respectively.

The original definition for CAI relative adaptedness values was that they should be based on a dataset of genes experimentally determined as being highly expressed (Sharp and Li 1987a). The fitness values built into CodonW were calculated using these criteria. CAI fitness values derived from genes identified solely based on a COA, are therefore not true CAI relative adaptedness values and as such should be treated with caution. The optimal codons recorded in the files “cbi.coa” and “fop.coa” were identified using a statistical test for significance which is dependent on sample size. Thus, the size of the sample taken from the extremes of the axis will affect the identified optimal codons. Despite these restrictions, this automatically discovered codon usage information can allow the reliable calculation of these indices in species for which there is limited experimental data see section 2.9.2.3 below.

2.9 Applications of CodonW

2.9.1 Codon usage indexes

2.9.1.1 The ENc-plot

Many species appear to have mutational biases, this is usually most apparent in the base composition of non-coding regions and synonymous codon positions. Such biases can cause non-random usage of codons (i.e. codon bias). However, the presence of codon bias is often mistakenly interpreted as evidence for selection, when the bias is an artefact produced by a mutational process. A useful index for examining the relationship between codon bias and mutation bias is the effective number of codons (EN_C) (Wright 1990). It is possible to predict the affect that any G+C bias will have on this index, assuming codon choice is solely a function of mutation bias (i.e. G+C = GC_{3s}). Equation 2-15 approximates the expected value of EN_C if codon bias is solely a function of GC_{3s} (Note: the original formulae, Equation 4 in Wright (1990), was misprinted).

Equation 2-15

$$EN_c = 2 + S + \left(\frac{29}{S^2 + (1-S)^2} \right)$$ S is the frequency of G+C (i.e. $GC3_S$)

Wright (1990) suggested the EN_c-plot (EN_C plotted against $GC3_S$ with Equation 2-15 superimposed) as part of a general strategy to investigate patterns of synonymous codon usage. Genes whose codon choice is constrained only by a G+C mutation bias, will lie on or just below the curve of the predicted values (Wright 1990). EN_c-plots of genes from six representative species (*Clostridium acetobutylicum, Streptomyces lividans, H. sapiens, E. coli, D. melanogaster* and *S. cerevisiae*) are presented in Figure 2-4. These demonstrate some of the commonest features of codon usage.

C. acetobutylicum and *S. lividans* have extremely biased G+C base compositions. *C. acetobutylicum* (low G+C) have a higher frequency of A or T ending codons, while *S. lividans* (high G+C) almost exclusively uses G or C ending codons, these codon choices are reflected by low values for EN_C (the lower the EN_C the more biased the codon composition). However, when these values are compared with those predicted under the supposition that codon bias is solely the result of mutation bias, the predicted values are not markedly different. In other words, the observed codon bias is most easily explained as a product of G+C mutation bias. This is typical for many species with biased G+C compositions where, if there is selection between synonymous codons, it does not appear to be able to overcome the influence of G+C mutational bias.

E. coli and *S. cerevisiae* are typical of species that display translational selection for codons, they are also the species where the correlation between tRNA concentration and choice of preferred codon was first established (Ikemura 1981a; Ikemura 1982). Genes with low frequencies of optimal codons (F_{op}) have much lower codon bias compared with those genes with a higher F_{op}. Genes with a low F_{op} have a codon bias which is similar to that predicted if mutational biases only influenced codon choice. It is interesting to note that in *E. coli* many of the genes with the lowest F_{op} have also the lowest GC_{3s}, the atypical base composition and codon usage of these genes suggests that they have been acquired by horizontal gene transfer. A striking feature of the EN_c-plots for these species and a strong indicator of translational selection is that within a relatively narrow range of $GC3_S$ there is a large variation in EN_c ranging from 61 to 26.3 for *E. coli* and 61 to 24.1 for *S. cerevisiae*, the correlation between the observed and “expected” EN_c is also weak (r=0.157 for *E. coli* and r=0.159 for *S. cerevisiae*).

The EN_C-plot for *H. sapiens* displays codon usage features typical of mammals, with a much wider range in $GC3_S$ content (0.25 - 0.95) and codon bias (28 - 61) than seen in prokaryotes. This variation in codon bias has been interpreted as evidence that selective pressure on

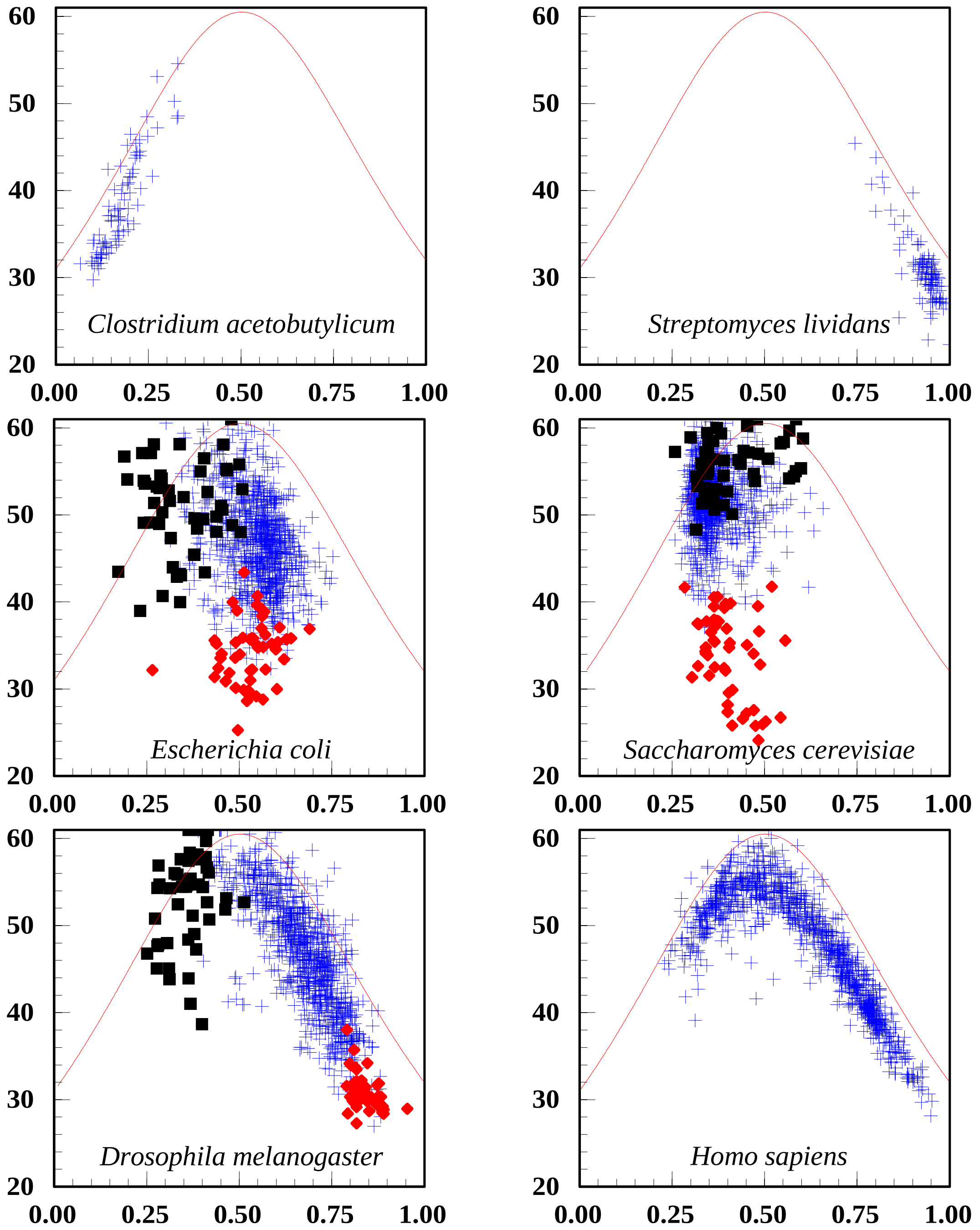

Figure 2-4. The effective number of codons (EN_c) plotted against GC_{3s}, for a representative sample of genes from *Clostridium acetobutylicum*, *Streptomyces lividans*, *E. coli*, *D. melanogaster, S. cerevisiae* and *H. sapiens*. The continuous curve plots the relationship between EN_c and GC_{3s} in the absence of selection. Genes shorter than 50 codons were excluded. A random sample of 1000 genes was plotted where the number of genes exceeds this figure. In species where optimal codon are know (i.e. *E. coli*, *D. melanogaster* and *S. cerevisiae*) the frequency of optimal codons (F_{op}) was used to categorise genes into the 5% with lowest (black box) and highest (red diamond) F_{op} values.

synonymous codons differ between genes (Mouchiroud and Gautier 1990; Newgard *et al*. 1986). The EN_C-plot summarises succinctly the relationship between $GC3_S$ and codon bias in *H. sapiens* where the variation in codon bias is most easily explained as an artefact of G+C mutational bias. This is typical of the general phenomenon of G+C variation with location (i.e. isochores) with a consequential variation in codon bias that has been observed in many vertebrate species (Ikemura and Wada 1991). There is little evidence for codon choice being shaped by selection for translation efficiency (Sharp *et al*. 1993).

Until the identification of translationally optimal codons in *D. melanogaster,* there was thought to be a dichotomy between the codon usage of unicellular and complex multicellular species (Shields *et al*. 1988). Although there do not appear to be isochores in *D. melanogaster* (Bernardi *et al*. 1985), *D. melanogaster* displays a wide range of variation in GC_{3s}. This is the consequence of selection for a subset of translationally optimal codons, which predominately contain G+C in their synonymous third position. Genes with the highest frequency of optimal codons have the lowest EN_c values and highest $GC3_S$ values, although superficially similar to the *H. sapiens* EN_c-plot a closer examination reveals subtle differences. There are more genes separate from the main gene cloud, there is also an increasing deviation from the expected bias curve with increasing codon bias. These in themselves are not enough to indicate selection for optimum translation. For further evidence it is necessary to examine the codon usage in more detail. The frequency of the codons CGU and GGU (both optimal codons) increase as codon bias increases, despite GC_{3s} also increasing. There is also a lack of correlation between GC_{3s} and the base composition of introns and surrounding non-coding sequence. Hence, while useful, an EN_C-plot in itself is not sufficient to distinguish between mutational biases and selection.

2.9.1.2 Is the codon usage of *S. cerevisiae* genes bimodal?

Using the codon bias index and cluster analysis of codon usage, the distribution of codon bias in *S. cerevisiae* has been previously described as bimodal (Sharp and Li 1987a; Sharp, Tuohy and Mosurski 1986). In a later analysis of 575 yeast genes, the original clear distinction between highly and lowly expressed genes was not as clear, but the distribution of codon bias remained “apparently bimodal” (Sharp and Cowe 1991). With the completion of the and CBI were calculated by CodonW for 6,218 annotated ORFs using the in-built *S. cerevisiae* codon usage data, the distribution of these indices can be examined in Figure 2-5. The indices were also calculated for a very long artificial sequence, whose codon usage had the same base composition bias as the *S. cerevisiae* genome (G+C = 40%). The values for this random sequence were 0.101, 0.382, and 0.022 for CAI, Fop, and CBI respectively.

Between 90% (CAI) and 94% (F_{op}) of the ORFs have index values equal to, or greater than, the index of the randomly generated sequence. This does not imply that ORFs, whose index value is less that for the randomly generated sequence, are under negative selection for optimal codons. Rather it implies that the vast majority of the ORFs have codon usage more biased towards an optimal codon choice than expected by chance alone. The average values for the indices CAI, F_{op}, and CBI were 0.174, 0.474, and 0.105 respectively. The correlation between the indices was very strong, the Pearson correlation coefficient between F_{op} and CBI was 0.99, and between CAI and F_{op} (or CBI) was 0.94. The distribution of CAI was the most leptokurtic and skewed. The indices CBI and F_{op} while having different means have similar kurtosis and skewness. The skewness is due to the strong bias for optimal codons, among a minority of genes. Although the precise cut-off is difficult to define, between 3% and 4% of the ORFs appear to be highly biased.

A comparison between these distributions and those in earlier yeast codon usage papers (Sharp and Cowe 1991; Sharp and Li 1987a), indicate that a disproportionate number of the early yeast sequences had highly biased codon usage. One of the principle reasons for this is that many of the highly biased genes are involved in metabolic pathways and are therefore easily identifiable by knockout mutagenesis. There is little evidence of codon bias having a bimodal distribution in yeast, rather the distribution is better considered as positively skewed. The original bimodal distribution was an artefact of the initially biased datasets.

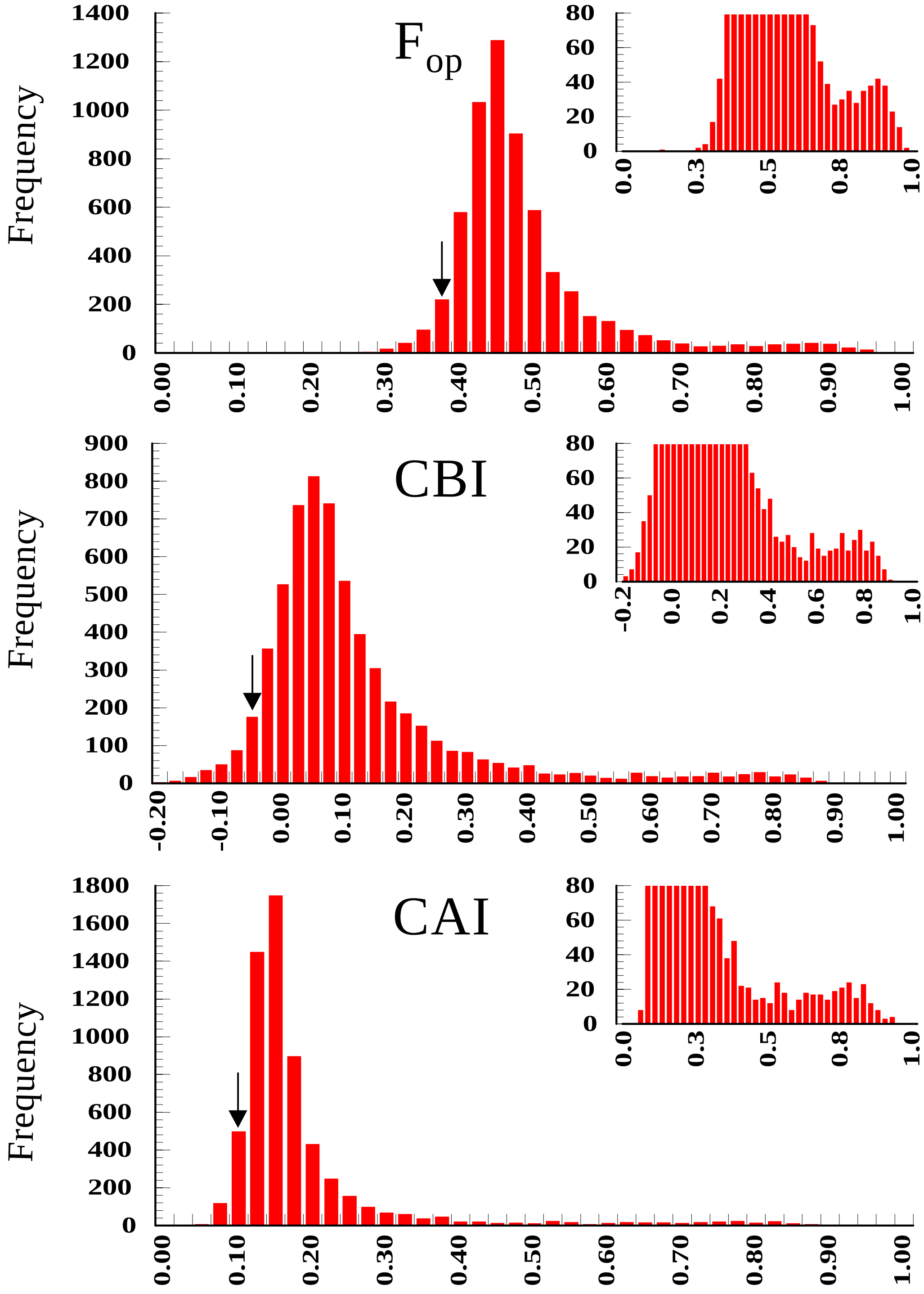

Figure 2-5. Distribution of three indices of codon usage calculated by CodonW for 6218 ORFs from the complete genome of *S. cerevisiae*. Histograms of the distribution of the Frequency of optimal codons (Fop), the codon bias index (CBI), and the codon adaptation index (CAI) are presented. The smaller inserted histogram displays the same distribution but with the vertical scale adjusted to enhance the distribution of the less frequent but more highly biased genes. The black arrow represents the index value for a sequence of almost infinite length with the same base composition as the *S. cerevisiae* genome (G+C = 0.40), see text.

2.9.2 Correspondence analysis

2.9.2.1 Correspondence Analysis of the Codon Usage of Cyanobacteria *Synechococcus sp.*

During the period of this thesis the whole genome sequence of the cyanobacteria *Synechocystis* sp. PCC6803 was completed by the Kazusa DNA Research Institute (Kaneko *et al*. 1996b), an analysis of the codon usage of this genome is presented in Chapter 3. The cyanobacteria *Synechococcus* sp. PCC6301 (2.7 Mb, G+C 55.1%) and *Synechococcus* sp. PCC7002 (2.7 Mb, G+C 49.1%) differ in both genome size and organisation from *Synechocystis* sp. PCC6803 (3.57 Mb, G+C 47.7%) (Kaneko *et al*. 1996a; Kotani and Tabata 1998). A sequence comparisons of *Synechococcus* sp. PCC7942 and *Synechocystis* sp. PCC6803, *gap1* and *gap2* genes found 40% and 25% respectively amino acid sequence divergence between the species.

Cyanobacterial species differ in morphology, physiology, biochemistry, and in genome structure. They include both filamentous and unicellular bacteria, with a photosynthetic apparatus of the higher plant type. (Tandeau de Marsac and Houmard 1987). The codon usage of unicellular cyanobacteria *Synechococcus* sp. have been analysed for relationships both between them and with other microbes and organelles, the codon usage was reported as being similar to that of lowly expressed *E. coli* genes (Krishnaswamy and Shanmugasundaram 1995). No selection for translationally optimal codons has been identified in cyanobacterial *Synechococcus* sp. It was unclear whether this was because this genus did not have selection for optimal codons or because their codon usage has never been subjected to a sufficiently sophisticated analysis. The majority of *Synechococcus* sequences have been submitted from three strains (PCC 7002, PCC 7942, and PCC 6301). These strains were originally classified as the separate species *Anacystis nidulans* (PCC 7942, and PCC 6301) and *Agmenellum quadruplicatum* (PCC 7002).

Sequences from *Synechococcus sp.* PCC 7002, *Synechococcus sp.* PCC 7942, and *Synechococcus sp.* PCC 6301 were extracted from GenBank (Release 95) using ACNUC the database query and retrieval program (Benson *et al*. 1994; Gouy *et al*. 1985). Those sequences shorter than 50 codons, sequences isolated from plasmids or transposable elements were removed. The final dataset contained 220 sequences, 55 from strain PCC 6301, 41 from strain PCC 7002, and 124 from strain PCC 7942.

The sequences were analysed using CodonW. A correspondence analysis of codon usage generated a first axis that explained 12.6% of the data inertia, with a second axis explaining 8.6% (see Figure 2-6). While the first axis did not explain a large amount of the variation it was correlated with EN_c, (r=0.66, p<0.001), while the second factor was correlated with GC (r=0.68, p<0.001) and GC_{3s} (r=0.645, p<0.001). A scatter plot of gene ordination on the principle axes indicates that the first factor is common to all the strains, while the second trend separated PCC 7002 from the other two strains (see Figure 2-7). CodonW identified putative optimal codons, using methodology, and assumptions discussed previously. Eleven genes were sampled from the extremes of principal axis and the putative optimal codons are presented in Table 2-6.

In total 18 putative optimal codons included the universally optimal codons (UUC, UAC, AUC, AAC GAC, and GGU). Two codons were identified for Leu, Arg, and Gly, while none were identified for Lys, Cys, and Val. The termination codon UAA is used preferentially in the higher bias dataset. In the high bias dataset 8 codons were identified as rare (i.e. RSCU < 0.10), these are indicated with a minus sign (-) in Table 2-6, these codons are also avoided in highly expressed *E. coli* genes (Sharp *et al*. 1992). The codons ACA, CGA and AGG were absent from the high bias dataset.

The 11 genes in the high bias group included 4 genes from Photosystem II (*psb*A1, *psb*A2, *psb*A3, and *psb*C), two chaperons (*gro*ES, *gro*EL), a gene from Photosystem I (*psa*E), the major and minor subunits of ribulose 1,5 biphosphatase, ribulose 1,5 biphosphatase carboxylase oxygenase, and a subunit of ATP synthase. The genes included in the lower biased subset were less well characterised, and included two membrane associated proteins, a phycocyanobilin lyase, an URF (unknown reading frame) upstream of a copper transport protein, and an URF downstream of glutamine synthetase.

The high bias dataset contains genes that are involved in anabolic and catabolic metabolic pathways, and which might be expected to be highly expressed. The set of putative optimal

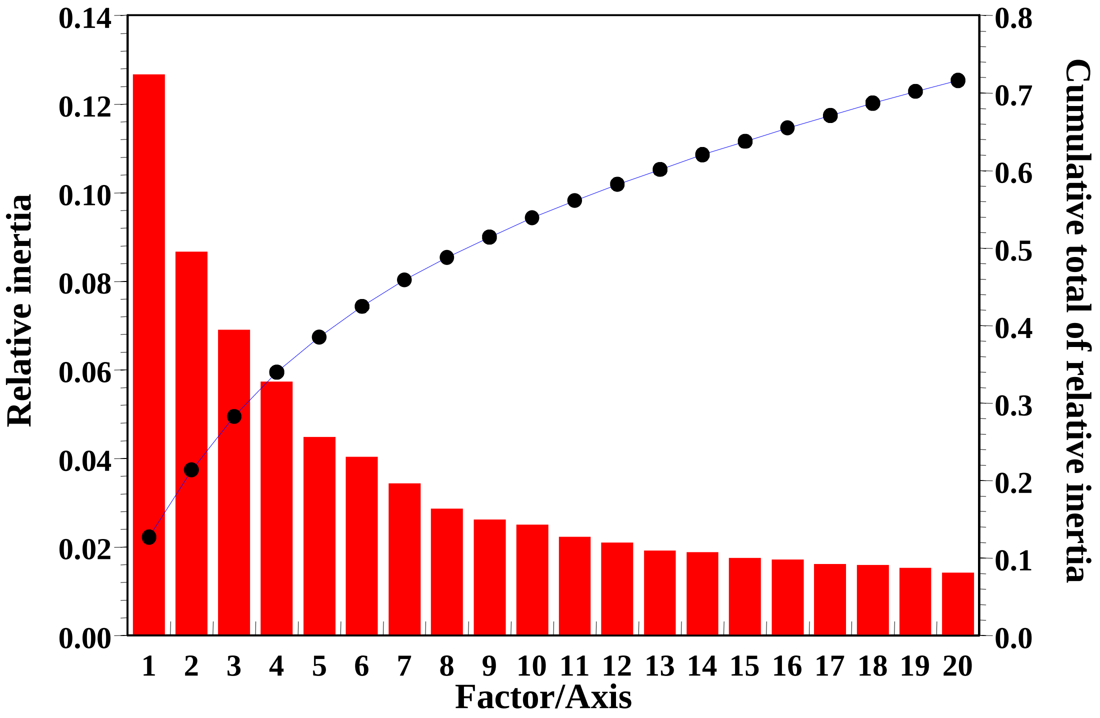

Figure 2-6. The relative and cumulative inertia, of the first 20 factors from a correspondence analysis of 220 genes from *Synechococcus sp.* codon usage.

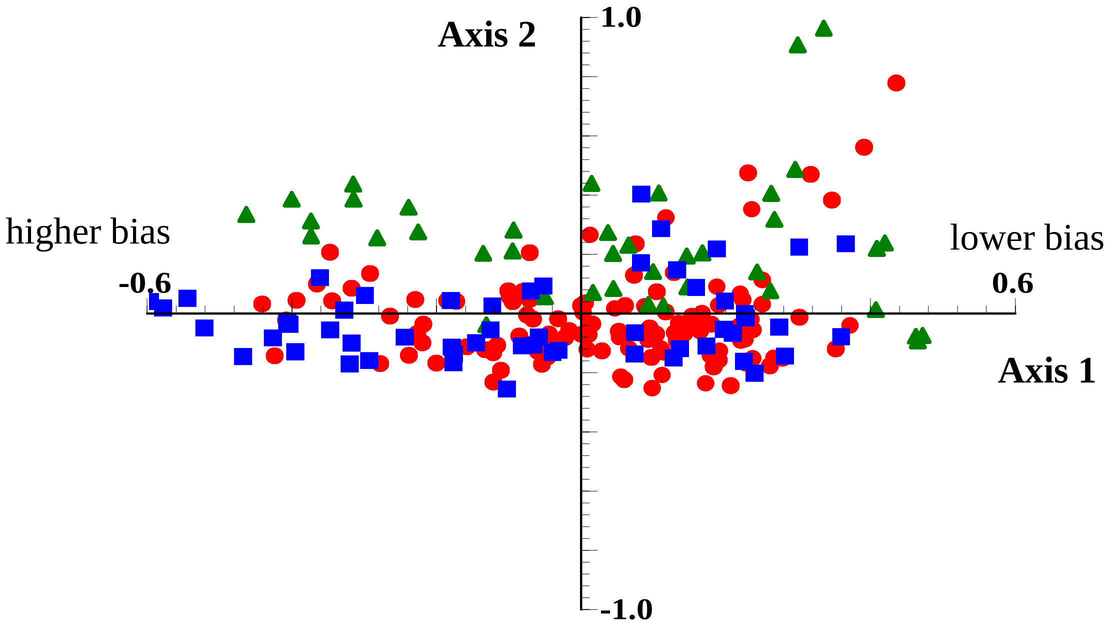

Figure 2-7 Ordination of 220 *Synechococcus sp.* sequences on the two principle correspondence analysis axes. These axes were generated by CodonW using the codon usage of the 59 synonymous codons. The strain number of each sequence is indicated by data marker used, PCC 7002 is indicated by a green triangle, PCC 7943 by a red circle, and PCC 6301 by a blue square.

		High Bias		Low Bias				High Bias		Low Bias	
		RSCU	CU	RSCU	CU			RSCU	CU	RSCU	CU
Phe	UUU	0.36	35	1.52	57	Ser	UCU	0.67	21	0.42	8
	UUC*	1.64	158	0.48	18		UCC	1.24	39	1.01	19
Leu	UUA	0.13	7	1.03	55		UCA-	0.10	3	0.69	13
	UUG	1.47	78	1.69	90		UCG@	2.17	68	1.49	28
	CUU	0.21	11	0.43	23	Pro	CCU	1.05	40	0.69	22
	CUC*	1.41	75	0.81	43		CCC*	1.65	63	1.01	32
	CUA-	0.00	0	0.79	42		CCA	0.13	5	0.72	23
	CUG*	2.78	148	1.24	66		CCG	1.18	45	1.57	50
Ile	AUU	0.50	36	1.56	69	Thr	ACU	0.71	36	0.49	16
	AUC*	2.50	180	1.24	55		ACC*	2.76	141	1.32	43
	AUA-	0.00	0	0.20	9		ACA-	0.00	0	0.71	23
Met	AUG	1.00	95	1.00	41		ACG	0.53	27	1.48	48
Val	GUU	1.16	81	1.00	36	Ala	GCU*	1.34	122	0.65	38
	GUC	1.16	81	0.89	32		GCC	1.05	96	1.14	67
	GUA	0.23	16	0.19	7		GCA@	0.99	90	0.70	41
	GUG	1.45	101	1.92	69		GCG	0.62	56	1.51	89
Tyr	UAU	0.28	15	1.15	47	Cys	UGU	0.76	14	0.65	13
	UAC*	1.72	91	0.85	35		UGC	1.24	23	1.35	27
Stop	UAA	2.45	9	1.09	4	Stop	UGA	0.00	0	1.09	4
Stop	UAG	0.55	2	0.82	3	Trp	UGG	1.00	67	1.00	41
His	CAU	0.15	6	1.42	59	Arg	CGU*	2.79	81	0.40	10
	CAC*	1.85	76	0.58	24		CGC*	2.55	74	1.33	33
Gln	CAA*	1.50	75	1.12	84		CGA-	0.00	0	1.25	31
	CAG	0.50	25	0.88	66		CGG	0.59	17	2.26	56
Asn	AAU	0.27	21	1.39	39	Ser	AGU	0.13	4	1.75	33
	AAC*	1.73	136	0.61	17		AGC*	1.69	53	0.64	12
Lys	AAA	1.23	86	1.04	35	Arg	AGA-	0.07	2	0.16	4
	AAG	0.77	54	0.96	32		AGG-	0.00	0	0.60	15
Asp	GAU	0.55	47	1.34	91	Gly	GGU*	1.97	169	0.86	40
	GAC*	1.45	125	0.66	45		GGC*	1.85	159	1.03	48
Glu	GAA*	1.25	142	0.86	53		GGA-	0.02	2	0.79	37
	GAG	0.75	85	1.14	70		GGG	0.15	13	1.33	62

Table 2-6 Correspondence analysis of 220 cyanobacteria *Synechococcus sp*. genes, generated a principal axis onto which the ordination of each gene was projected. The codon usage of eleven genes (5% of the total number of genes) from the extremes of the principal were pooled. The codon usage of both pools was compared using a two-way Chi squared contingency test, to identify optimal codons. For the purposes of this test dataset with the lower EN$_c$ were putatively assigned as highly expressed. The codon usage and RSCU of both datasets is shown. Those codons that occur significantly more often (p<0.01) are indicated with a (*), codons with p<0.05 are indicated with @, codons with a RSCU less than 0.10 are considered rare and are indicated with a minus ′-′. Number of codons in high bias dataset 3626 and in the low bias dataset 2442.\

codons include a set of universal optimal codons. The termination codon usage and rare codon usage is similar to highly expressed *E. coli* genes. Taken together this is strong evidence for selection of a preferred subset of codons in the cyanobacterial *Synechococcus sp.*, this selection appears similar to that seen in *E. coli* and is presumably for optimum translation.

2.9.2.2 Correspondence analysis of the amino-acid usage of *Salmonella typhimurium*

Codon usage diverges with increasing evolutionary distance, but in the closely related species *E. coli* and *S. typhimurium* it is very similar. Ikemura (1985) demonstrated that in *E. coli, S. typhimurium, S. marcescens,* and *S. cerevisiae* codon bias was correlated with the abundance of the cognate tRNA. In *E. coli* and *S. typhimurium*, homologous genes have very similar codon bias (Sharp 1991; Sharp and Li 1987b). Either there has been a insufficient number of substitutions for a difference to be detected, or selection pressure on synonymous codon choices has remained roughly the same in the two organisms since they diverged 10^8 years ago (Ochman and Wilson 1987a). There is a very strong negative correlation between the substitution rate and codon bias (Sharp 1991).

In contrast to codon usage, the interspecific variability of amino-acid usage is low (Grantham, Gautier and Gouy 1980a). However, a correspondence analysis of *E. coli* amino-acid usage found a strong positive correlation between amino-acid composition and codon bias. This has been shown, after hydrophobicity, to be the second strongest trend in the amino-acid composition of *E. coli* (Lobry and Gautier 1994). Surprisingly this is a more significant trend than aromaticity, amino-acid volume, or charge (Lobry and Gautier 1994).

To investigate whether the same trends in amino-acid usage were present in *S. typhimurium*, the amino acid usage of a dataset of 503 genes was investigated. This dataset was a subset of all *S. typhimurium* coding sequences in Release 59 of the EMBL database (Emmert *et al*. 1994) which contained cross-references to the SwissProt protein database (Bairoch and Boeckmann 1994). This query was constructed using the command line interface of the SRS sequence retrieval system (Etzold and Argos 1993). Sequences shorter than 50 codons, associated with plasmids or transposable elements, poorly documented open reading frames, and partial sequences were

removed. The sequences were then conceptually translated by CodonW into protein and then analysed using the "correspondence analysis of amino-acids" option of CodonW.

The first three axes generated of the correspondence analysis explained 40% of the amino-acid variation. The position of each protein on the first and third axis are shown in Figure 2-8a. The first axis accounted for 16% of the variability in amino-acid usage. This axis is highly correlated (r=0.88, $p < 10^{-4}$) with the hydrophobicity GRAVY score (Kyte and Doolittle 1982), see Figure 2-8b. These figures are similar to those reported for *E. coli* where the principle factor explained 17% of the variation and was highly correlated (r=0.90, $p < 10^{-4}$) with the GRAVY score. The second axis accounted for slightly more of the variation in *S. typhimurium* (13.8%) than in *E. coli* (13%). In *E. coli* this axis was correlated (r=0.55, $p < 10^{-4}$) with CAI, in *S. typhimurium* the correlation with CAI is a weaker (r=0.30, $p < 10^{-3}$), see Figure 2-8d, CAI has a better correlation with axis 1 (r=0.33, $p < 10^{-3}$), see Figure 2-8e. In *S. typhimurium* the third axis accounted for 9.4% of the variation and was correlated with aromaticity (r = 0.64, $p < 10^{-4}$), this is similar to the values reported for *E. coli* where it explained 10% of the variation and it too was correlated (r=0.70, $p < 10^{-4}$) with aromaticity (Lobry and Gautier 1994), see Figure 2-8c.

The most important trend in the amino-acid usage of *S. typhimurium* genes is the usage of hydrophobic amino-acids, the third most important trend is the usage of aromatic amino-acids. These trends are common with those found in *E. coli* (Lobry and Gautier 1994). However, unlike in *E. coli* where the second most important COA trend is strongly correlated with CAI, in *S. typhimurium* this correlation is much weaker. It is not clear if the weak correlation with CAI reflects a less effective selection of amino acids for translational efficiency or if it reflects the high degree of sequence similarity between these two species.

2.9.2.3 Automated identification of optimal codons in genomes.

The genome sequencing projects have resulted in the situation where the number of characterised genes for some species (e.g. *Mycoplasma genitalium*, *Haemophilus influenzae,*

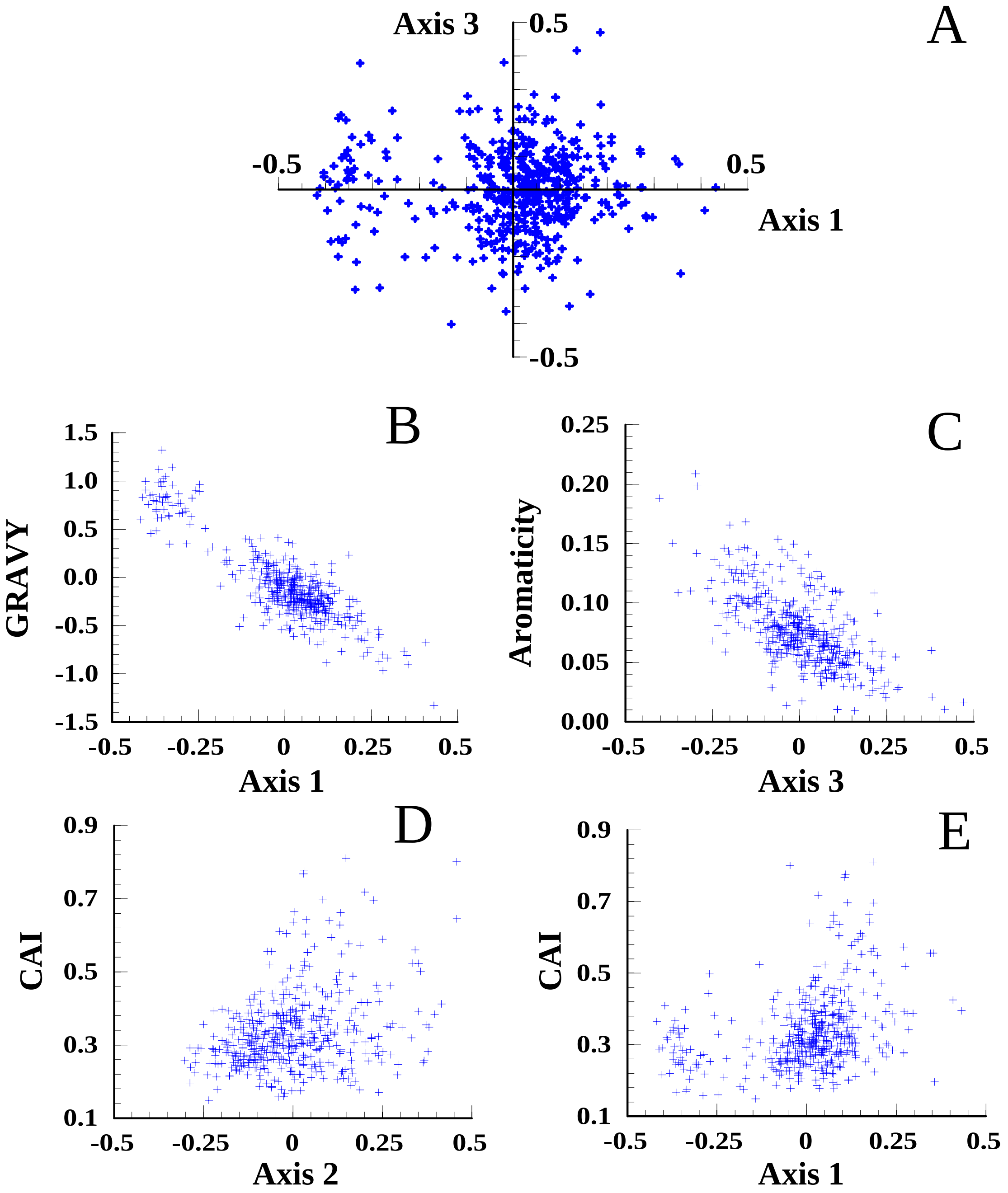

Figure 2-8 (A) Correspondence analysis of amino-acid usage in 503 *S. typhimurium* genes. Each point is a gene plotted at its coordinates on the first and third axes, the first axis is on the horizontal (B) Correlation of the global hydrophobicity of proteins (GRAVY score) with the COA axis 1 (C) Correlation of aromaticity with COA axis 3.(D) Correlation between axis 2 and CAI (E) Correlation between axis 1 and CAI

or no experimental validation of gene products, regulation, or translation levels. To investigate the codon usage of these new genomes, it will be necessary to apply many codon usage analysis tools blindly, with little or no experimental evidence. As the number of sequenced genomes increases this will become an increasing problem. A potential problem that codon usage analyses may encounter, is that many of the sequences identified by these genome projects, have been automatically identified as potentially protein coding by programs that, at least in part, base their decisions on codon usage information. This will result in false ORFs having codon usage patterns that are similar to real genes, this could potentially confuse codon usage analysis programs.

We have seen above (see section 2.9.1.2) that the initial investigation of codon usage and identification of optimal codons in *S. cerevisiae* were based on datasets that overrepresented highly expressed genes (Ikemura 1982; Sharp and Cowe 1991; Sharp and Li 1987a). This raises the question, if presented with a more representative sample of *S. cerevisiae* genes, or if presented with this dataset with no biochemical information, could the same optimal codons still be identified? As CodonW was designed to allow the rapid investigation of codon usage and to facilitate the identification of optimal codons, to answer this question CodonW was used to analyse a dataset of 6203 *S. cerevisiae* S288C genes. This dataset is a subset of the 6218 gene dataset used in section 2.9.1.1 but with genes shorter in length than 50 codons removed to reduce signal noise.

The default CodonW correspondence analysis parameters were used for the COA of the dataset codon usage. The number of genes initially sampled from both extremes of the principle axis was the default 5% or genes (i.e. 310 genes). This allowed 24 codons to be putatively identified as optimal. These included the 22 codons previously identified as being optimal by Sharp and Cowe (1991), and the codon GCC identified as optimal by Bennetzen and Hall (1982). The codon CGU has not previously been identified as being optimal in *S. cerevisiae*, but this codon was just significant (χ^2 = 6.646). Its RSCU (0.86) in the high bias dataset was not particularly high but it was significantly more than in the low bias dataset (RSCU = 0.63).

The optimal codons were automatically recorded to files in the format required by CodonW to calculate the codon usage indices F_{op} and CBI using a personal choice of optimal codons. Relative adaptedness (ω) values, as used by the CAI index, were also recorded. These ω values were calculated from the codon usage of the high bias group of genes. Each index was calculated using the codon usage information "discovered" by CodonW, and the index values compared with values calculated using the *S. cerevisiae* codon usage information built into CodonW (Sharp and Cowe 1991; Sharp and Li 1987a).

Results from regression analyses of the correlations between the new and original index values are shown in Table 2-7. As expected with 22 optimal codons in common, the correlation between the new and original F_{op} ($r^2 = 0.989$) and CBI ($r^2 = 0.993$) values are very strong. The axis intercept values for both indices are close to zero, but the slopes are greater than 1. The increased number of optimal codons causes the newer index values to be on average slightly higher. The correlation for CAI ($r^2 = 0.963$) is lower than the correlations of other two indices. A plot (see Figure 2-9) of the correlation reveals that the relationship between the new and original CAI values is non-linear.

The optimal codons, and therefore F_{op} and CBI, are more robust to changes in the codon usage of groups used to identify them. CAI is calculated from ω values and is more sensitive to changes in the codon usage composition of the group used to calculate them. The original *S. cerevisiae* ω values were calculated from the codon usage of only 23 highly expressed genes (Sharp and Li 1987a). Previously in this chapter, I have shown that only between 3 % and 4% of ORFs from this dataset are highly biased, therefore the default 5% cut-off includes many genes that are not highly expressed.

The correspondence analysis was repeated, with the number of genes sampled from the extreme of the principle axis limited to 100. The optimal codons identified were the same as for the previous analysis, with the exception of CGU, which was no longer optimal. This improved the correlation, intercept and slope of F_{op} ($r^2 = 0.993$) and CBI ($r^2 = 0.995$), by increasing the similarity of the set of optimal codons. The correlation between the new and original CAI ($r^2 = 0.997$) increased dramatically, see Figure 2-9. The slope is 1.029, indicating the old and new values for CAI are not identical, however the rank

Index	Group	b0	b1	r2
CBI	5%	0.0073	1.0847	0.993
	100	0.0041	1.0595	0.995
Fop	5%	-0.0081	1.0610	0.989
	100	-0.0098	1.0502	0.993
CAI	5%	-0.1602	0.9849	0.963
	100	0.0106	1.0294	0.997

Table 2-7 Linear regression of the correlation between the CBI, F_{op} and CAI indices calculated for 6203 yeast ORFs using the *S. cerevisiae* codon usage information in-built into CodonW, with the values for the indices calculated using codon usage information automatically discovered by CodonW by a correspondence analysis of the yeast ORFs' codon usage. Optimal codons were identified by a two-way Chi-squared contingency test, between groups taken form both extremes of axis 1. CAI was calculated using relative adaptedness (ω) values calculated from a group highly biased (expressed) genes. The number of genes in each group was either 5% (the default) of the total number of genes (i.e. 310 genes) or 100 genes. The parameter b0 is the axis intercept, b1 is the slope, and r^2 is the squared correlation coefficient.

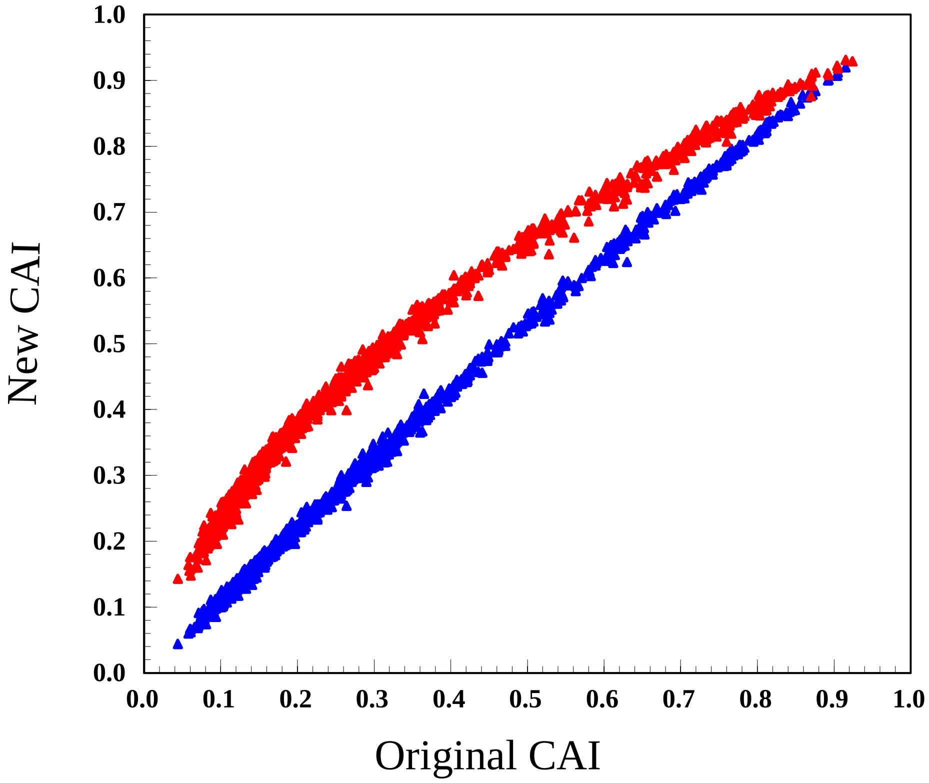

Figure 2-9 Correlation between CAI index values calculated for 6203 *S. cerevisiae* ORFs. The CAI index was calculated by CodonW using either the original *S. cerevisiae* relative adaptedness (ω) values in-built into CodonW or new ω values automatically calculated from the codon usage of two groups of putatively highly expressed genes. The composition of these two groups was determined by gene ordination on axis 1 of a codon usage COA. Either the first 5% of genes or the first 100 genes based were selected, the CAI indices calculated using ω values from the first 5% of genes are indicated by red triangles, while those calculated using the first 100 genes are indicated by blue triangles.

order of the new and original CAI values are almost the same. Such a strong correlation was surprising as the original ω values were calculated from 23 genes experimentally determined as being highly expressed, while the newer ω values were calculated from the codon usage of 100 genes automatically selected by CodonW.

These results demonstrate that CodonW was able to accurately identify optimal codons from a correspondence analysis of the codon usage of ORFs from the entire genome gene compliment. The codon usage indices F_{op} and CBI were less sensitive to the number of genes used to identify optimal codons, or to changes in the set of optimal codons. When analysing the codon usage of the output of a genome-sequencing project, 5% is too large a sample for calculating ω values or the identification of optimal codons, probably 1% of total gene compliment is more practical. CodonW can identify optimal codons and estimate CAI ω values with surprising accuracy based simply on patterns of codon usage.

3 Codon usage *in Lactococcus lactis*

3.1 Codon usage in the Gram-positives

The majority of the prokaryotic genes analysed in the early codon usage papers were derived from the Gram-negative proteobacteria (Gouy and Gautier 1982; Grantham, Gautier and Gouy 1980a; Grantham *et al*. 1981; Grantham *et al*. 1980b). This was not caused by a bias in the selection of available sequences but rather reflected the bias in the genetics (including gene sequencing) towards Gram-negative proteobacteria. A consequence of this is that our understanding of codon usage in Gram-negative prokaryotes, in particular *E. coli*, is more advanced than for Gram-positives. As interest in the genetics of the Gram-positives increased, there was a corresponding increase in the number of available sequences in the databanks. Yet, despite this, our knowledge of codon usage variation in these species remains at best fragmentary. The Gram-positive species that has been most extensively studied is the Low G+C species *Bacillus subtilis.* The codon usage bias of *B. subtilis,* while less pronounced than that of *E. coli,* has been shown to be correlated with gene expression level (Shields and Sharp 1987). This correlation has been ascribed to selection for a subset of translationally optimal codons in highly expressed genes (Shields and Sharp 1987).

A principal coordinate analysis of the codon usage of genes from *Lactobacillus* species, another Low G+C Gram-positive prokaryote identified a subset of codons, which have been described as universally optimal (Sharp and Devine 1989) and which were used preferentially by the most highly expressed *Lactobacillus* genes (Pouwels and Leunissen 1994). Apart from this common theme in codon usage there was also a remarkable divergence in codon usage patterns between *Lactobacillus* species, reflecting the phylogenetic divergence of this genus (Pouwels and Leunissen 1994).

Codon usage in *Clostridium acetobutylicum,* another Low G+C Gram-positive prokaryote displays a strong bias (Croux and Garcia 1992; Winkler and Wood 1988), however the bias is similar in genes expressed at a range of expression levels. This bias appears to be a consequence of an overall genome mutational bias for A+T nucleotides resulting in a bias for A+T ending codons. Likewise, if there is selection for translationally optimal codons in the

High G+C Gram-positive *Streptomyces* species it is masked by mutational biases (Ohama *et al*. 1990; Ueda *et al*. 1993; Wright and Bibb 1992).

The codon usage of ORFs from the High G+C Gram-positive *Nocardia lactamdurans* and *Micrococcus luteus* species again display a lack of variation in codon usage but with an overall bias towards G+C ending codons (Ohama, Muto and Osawa 1990). However, a correspondence analysis of codon usage in the High G+C acid-fast Gram-positive *Mycobacterium tuberculosis* did identify variation in codon usage between genes (Andersson and Sharp 1996b). The primary source of this variation was the usage of a subset of codons, which were tentatively identified as translationally optimal. The variation was attributed to selection for optimal codons genes with higher expression levels. A systematic variation in the degree of codon usage bias has also been reported for cosmid ORFs from the High G+C *Mycobacterium leprae* genome sequencing project (Fsihi and Cole 1995). An analysis of synonymous codon usage, in the less G+C content biased but related species *Corynebacterium glutamicum*, indicates an increased codon usage bias in highly expressed genes (Eikmanns 1992; Malumbres, Gil and Martin 1993). Both identified what they described as optimal *C. glutamicum* codons using two similar methods, but based on the overall usage of codons rather than by examining the codon usage of highly expressed genes.

3.2 L. lactis - an overview

The industrial fermentation of milk by lactic acid bacteria represents one of the largest engineered ecosystems in the world; it has been estimated that per year there are 10^{23} bacteria involved in these bioconversions (de Vos and Vaughan 1994). The lactic acid bacteria include strains of the mesophilic *L. lactis, Lactobacillus casei* and *Leuconostoc lactis* and the thermophilic *Streptococcus thermophilus*, *Lactobacillus helveticus* and *Lactobacillus bulgaricus*. Of these bacteria, the best-studied species in terms of lactic acid synthesis is *L. lactis*, which is used to ferment lactose to lactic acid during cheese manufacture.

The genus *Lactococcus,* originally assigned to the genus *Streptococcus*, is a distinct and homogenous group of Low G+C content lactic acid bacteria (Collins *et al*. 1989; Olsen *et al*. 1994). However, within the genus the status of *L. lactis* subsp. *lactis* (*Streptococcus lactis*) and *L. lactis* subsp. *cremoris* (*Streptococcus cremoris*) as separate species or subspecies have been the subject of considerable debate (Delorme *et al*. 1994). Similarities between the phenotypes and DNA hybridization studies, which found 80% similarity for the type strains (Jarvis and Jarvis 1981), prompted Garvie & Farrow (1982) to suggest that *S. cremoris* and *S. lactis* belonged to a single species. They were later reassigned to the genus *Lactococcus* (Schleifer *et al*. 1985). This degree of homology can be contrasted with the similarity between *E. coli* strains (1-2% divergence) and the 16% average divergence between *E. coli* and *S. typhimurium* sequences (Sharp *et al*. 1995a). The subspecies classification of these strains was supported by a high degree of similarity in the sequence of their 16S rRNA genes (Collins *et al*. 1989). Other investigations found that some genes only weakly hybridised between the species (Godon *et al*. 1992). This observation was expanded to show that there were large regions of mosaic structure in the chromosome of *L. lactis* subsp. *cremoris* (Delorme *et al*. 1994). Divergence was low (10-20 %) in the conserved regions while in variable regions divergence was much higher (35-36 %). The divergence did not appear to be the result of rapid genetic drift, and the variable regions were characterised by a higher G+C content than is typical for genes for the genus *Lactococcus*. Unusual codon bias and the presence of a repeated DNA element, also suggested this divergent region had been acquired by horizontal transfer (Delorme *et al*. 1994).

L. lactis is of major economic importance in the dairy industry, and has a GRAS (generally recognised as safe) designation. This has generated a significant interest in the biochemistry, molecular biology and genetics of these strains (van de Guchte *et al*. 1992). The uptake of lactose and its conversion to lactic acid are essential for the normal growth of *L. lactis* in milk since this process constitutes the major metabolic pathway for its energy production (Llanos *et al*. 1992). This conversion requires the interplay of three metabolic pathways: the phosphoenolpyruvate-dependent phosphotransferase transferase system, the tagatose-6-phosphate pathway, and the Embden-Meyerhoff pathway (EMP) (Thompson 1987). Lactose is brought into the cell by the lactose transport system (de Vos and Vaughan 1994) and is then converted by phospho-β-galactosidase to glucose and galactose 6-phosphate. The glucose is then metabolised along the EMP, and galactose 6-phosphate is converted via the tagatose-6-phosphate pathway to triose phosphates, which in turn enter the EMP (Llanos, Hillier and Davidson 1992) with a net production of two ATP molecules per molecule of hexose consumed. The genes encoding the EMP are thought to be chromosomally located, as plasmid-cured strains of *L. lactis* subsp. *lactis* retain the ability to ferment lactose (Efstathiou and McKay 1976).

A characteristic of many lactic acid producing bacteria, including *L. lactis*, is the production of bacteriocins (Klaenhammer 1993; Sahl 1994). Nisin produced by some strains of *L. lactis*, subsp. *lactis* and a permitted food additive in the United Kingdom, is the most highly characterised bacteriocin produced by lactic acid bacteria (Harris *et al*. 1992). The genes encoding bacteriocins, such as lactococcin, are commonly located on plasmids or transposable elements and can be transferred horizontally between species (Altay *et al*. 1994; Dufour *et al*. 1991; Gireesh *et al*. 1992; Prevots *et al*. 1994).

3.3 Codon preference in *L. lactis*

While relatively little is known about codon usage in *L. lactis*, there has been a considerable amount of interest in the codon usage of this species. As it is fermented on such a large scale it seems likely that its effective population would be large enough for translational selection between codons to overcome mutational drift. The codon usage of the *L. lactis* subsp. *lactis pbg* gene and *L. lactis* subsp. *cremoris* protease genes were the first to be reported (Kok *et al*. 1988; Porter and Chassy 1988). Van de Guchte *et al.* (1992) tabulated the codon usage of 23 *L. lactis* genes, although not all of these genes were chromosomally located and included a putative transposase of IS904. They observed that codon usage could vary markedly between genes. The codon usage of amino-acid biosynthesis genes was examined for evidence that codon usage could be a mechanism for the regulation of gene expression (Chopin 1993). The codon usage of the *Lactococcus lactis* lactic acid synthesis (*las*) operon which encodes phosphofructokinase (*pfk*), pyruvate kinase (*pyk*), and lactate dehydrogenase (*ldh*) has been reported as being markedly more biased than the codon usage of 27 other chromosomally located *L. lactis* genes. This has been used to argue that the genes from the *las* operon are highly expressed (Llanos *et al*. 1993). The codon usage of glyceraldehyde-3-phosphate dehydrogenase (*gap*) and triosephosphate isomerase genes (*tpi*) was also compared to 84 *L. lactis* chromosomal genes and ORFs, and a bias in the codon usage of these glycolytic genes was also reported (Cancilla *et al*. 1995a; Cancilla, Hillier and Davidson 1995b).

Despite this interest in *L. lactis* codon usage none of these analyses reached significant conclusions in relation to codon usage apart from noting that there was variation in codon bias between genes. Identification of optimal codons was not attempted although the avoidance of a subset of codons was noted. The lack of conclusions was due, at least in part, to the small sample sizes and a reliance on simple tabulation of codon frequencies and the masking effects of mutational biases towards A+T rich codons.

3.4 Analysis of codon usage

3.4.1 Methods

Sequences: DNA sequences were extracted from the sequence database GenBank 95 using the retrieval program ACNUC (Gouy *et al*. 1985). To reduce sample noise, partial sequences and reading frames of less than 50 codons were removed (unless otherwise noted). Genes that were good candidates for having been horizontally transferred, i.e. transposable elements, and plasmid genes were also removed.

Redundancy: Duplicate entries in the datasets were identified by using the BLAST (Altschul *et al*. 1990) suite of programs to search each gene in the dataset for homology with all other genes within the same dataset. Genes were considered as duplicates if any pair had greater than 98% sequence identity. These genes were removed unless there was evidence that they were paralogous.

Identification of homologous genes: Genes which were homologous between species were initially identified as putative homologues by protein sequence similarity searches using BLASTP (Altschul *et al*. 1990) to search the SwissProt (Bairoch and Boeckmann 1994), PIR (George *et al*. 1994) and TREMBL (Bairoch and Apweiler 1998) protein sequence databases. These putative gene pairs were then used as the basis for a literature search. Those sequences that could be identified in the literature as homologous were recorded as such. Sequence pairs that had high sequence similarity (>50% amino acid identity) but where function had not been ascribed to both pairs, were labelled as being identified by similarity only.

Identification of Open Reading Frames: All open reading frames (ORF) with unknown function were searched against a non-redundant dataset of all known protein coding sequences and all putative coding ORFs, this dataset was a nonredundant composite of SwissProt, PIR and TREMBL. Protein sequence similarity was identified using BLASTP. Unidentified ORFs that displayed significant sequence similarity ($p<10^{-5}$) to a previously identified gene were putatively identified on the basis of that homology. Where *L. lactis* ORFs had significant similarity ($p<10^{-5}$) to an unidentified ORF in another species, this similarity was also noted. Where an ORF did not display significant similarity to any other coding sequence, it was labelled as an Unidentified Reading Frame (URF).

Analyses: All calculations of codon usage, codon usage indices and correspondence analyses were generated using the program CodonW. Explanations of the codon usage indices, correspondence analyses and methods calculated by CodonW have already been discussed (see Chapter 2). Rates of synonymous and non-synonymous substitution were estimated using the method of Li *et al.* (1985) as amended by Li (1993). Multiple sequence alignments were generated by Clustal W (version 1.6) (Thompson *et al*. 1994).

3.4.2 Results

3.4.2.1 Codon usage in Lactococcus lactis

The primary dataset used to analyse codon usage was a subset of *L. lactis* protein encoding reading frames. As discussed above, *L. lactis* is recognised to consist of at least two subspecies *L. lactis* subsp. *cremoris* and *L. lactis* subsp. *lactis.* However, the majority of genes have been sequenced from untyped strains and previous reports based on DNA hybridisation have found that strains have been assigned to the wrong subspecies (Godon *et al*. 1992). Therefore, in these analyses, both subspecies are considered as a single species. The His operon from *L. lactis* subsp. *cremoris* has been previously identified to have anomalous G+C content and has been postulated to have been acquired by horizontal transfer (Delorme *et al*. 1994) For this reason the His operon from *L. lactis* subsp. *lactis* rather than that from *L. lactis* subsp. *cremoris* was used in these analyses.

In GenBank release 95, there were 274 sequence entries from *L. lactis* species. These contained 465 putative coding genes, which included 7 short ORFs (< 50 codons). These short ORFs, which included the ribosomal gene *rpm*G (length = 48 AA) were removed. Although excluded from the COA (correspondence analysis), *rpmG* was included in subsequent analyses. Partial sequences were removed, with the exception of 12 genes that were no more than 15% truncated. These 12 partial ORFs included *rpm*O, *uhp*T, *clpA*, and *rps*M. The first ORF from GenBank accession no. X62621 that was annotated as 361 bp was truncated to the correct 360 bp ORF.

There was a high level of redundancy in the dataset of 465 putative ORFs, for example there were 16 copies of ldh, 3 copies of the lac operon. The *L. lactis lac* operon is frequently plasmid-borne and it is believed to undergo horizontal transfer between species, therefore *lac* operon genes were excluded, as were all plasmid (196) and IS (27) associated genes. Lactococcin, bacteriocin, and nisin (47) genes which are frequently located in transposable elements and on plasmids (Romero and Klaenhammer 1993) were also excluded from the initial dataset. This series of purification steps reduced the original 465 ORFs to 124 ORFs.

When this initial dataset of chromosomal genes was analysed by correspondence analysis and using codon usage indices, the *L. lactis* para-aminobenzoic acid synthetase gene *pabB* (Arhin and Vining 1993); (GenBank accession no. M64860) was found to have very atypical codon usage. These initial results, later confirmed, suggested that this gene was not native to the *L. lactis* genome (discussed in more detail in section 3.4.2.6). The gene *pabB* was therefore removed from the *L. lactis* dataset. Another gene with unusual codon bias was the glutaredoxin-like protein *nrd*H, however, this gene was short (72 codons) and analysis of the contribution of inertia to each axis indicated that it did not significantly influence the first four axes, therefore it was not excluded. This reduced the main dataset to 123 ORFs. A tabulation of the total codon usage of this dataset is presented in Table 3-1. The overall genomic G+C content of *L. lactis* is estimated to be between 37-41% (Sneath 1986). This reflects a mutational bias towards A+T which in the absence of other selection pressures, would be expected to increase the RSCU values for synonymous A+U ending codons to greater than 1. The codon usage presented in Table 3-1

		N	RSCU			N	RSCU			N	RSCU			N	RSCU
Phe	UUU	1446	1.49	Ser	UCU	664	1.50	Tyr	UAU	1036	1.48	Cys	UGU	145	1.57
	UUC	498	0.51		UCC	127	0.29		UAC	365	0.52		UGC	40	0.43
Leu	UUA	1228	1.79		UCA	935	2.12	TER	UAA	82	2.10	TER	UGA	26	0.67
	UUG	909	1.33		UCG	139	0.31		UAG	9	0.23	Trp	UGG	350	----
	CUU	1179	1.72	Pro	CCU	503	1.39	His	CAU	551	1.42	Arg	CGU	645	2.74
	CUC	333	0.49		CCC	103	0.28		CAC	227	0.58		CGC	143	0.61
	CUA	260	0.38		CCA	738	2.04	Gln	CAA	1382	1.74		CGA	186	0.79
	CUG	196	0.29		CCG	103	0.28		CAG	204	0.26		CGG	85	0.36
Ile	AUU	2229	2.07	Thr	ACU	874	1.49	Asn	AAU	1616	1.55	Ser	AGU	545	1.23
	AUC	712	0.66		ACC	307	0.52		AAC	470	0.45		AGC	240	0.54
	AUA	292	0.27		ACA	921	1.57	Lys	AAA	2622	1.69	Arg	AGA	299	1.27
Met	AUG	1022	----		ACG	245	0.42		AAG	475	0.31		AGG	52	0.22
Val	GUU	1471	2.11	Ala	GCU	1402	1.69	Asp	GAU	1735	1.45	Gly	GGU	1171	1.59
	GUC	475	0.68		GCC	474	0.57		GAC	661	0.55		GGC	367	0.50
	GUA	518	0.74		GCA	1127	1.36	Glu	GAA	2567	1.71		GGA	1084	1.47
	GUG	329	0.47		GCG	322	0.39		GAG	444	0.29		GGG	323	0.44

Table 3-1. Codon usage in *L. lactis* Codon frequency tabulation of 123 *Lactococcus lactis* genes and ORFs. N is codon frequency, RSCU is relative synonymous codon usage. Total number of codons in table is 42228 codons.

displays such an excess of A or U ending codons. However, three adenine ending codons CUA, AUA, and GUA are used far less frequently than expected, with RSCU values of 0.38, 0.27, and 0.74 respectively. The scarcity of these three codons is not simply due to the avoidance of the dinucleotide UA. The most frequent Leucine codon is UUA, although the UUA codon is avoided in highly expressed genes. Despite the scarcity of these three UA ending codons, UA is the sixth most frequent dinucleotide. UA ending codons are used less frequency than expected in many prokaryotic species including A+T rich species (Winkler and Wood 1988). In *E. coli* and *S. typhimurium,* CUA (Leu) and AUA (Ile) are each recognised by a single relatively rare tRNA (Ikemura and Ozeki 1982). If this also holds for *L. lactis,* it could explain the scarcity of these codons, but not the scarcity of GUA. In *E. coli* and *S. typhimurium* GUA is a common codon and is recognised by the relatively abundant $\mathrm{tRNA}_1^{\mathrm{Val}}$. However, $\mathrm{tRNA}_1^{\mathrm{Val}}$ also recognises GUU as does the less abundant $\mathrm{tRNA}_2^{\mathrm{Val}}$ and according to *rule 5* proposed by Ikemura and Ozeki (1982), “where a codon is recognised by two or more tRNAs it will be used more frequently than a synonym recognised by a single tRNA”.

3.4.2.2 Correspondence analysis

The theory and practice of Correspondence analysis (COA) has been extensively discussed in previous chapters. A COA of the codon usage (CU) of the initial dataset composed of 123 *L. lactis* chromosomal genes yielded a first axis that explained 20.0% of the total variation in this CU data. The variation explained by the second axis (7.5%) was approximately one third of the variation explained by the first axis, with each subsequent axis explaining a decreasing amount of the variation (see Figure 3-1). In some species where the major influence on codon choice is selection for optimal translation the first axis can explain between 30% and 40% of the total variation (Andersson and Sharp 1996b). Nevertheless, this sharp reduction in the variation explained by the first and subsequent axes is often indicative of a single trend in the systematic variation among genes. The ordination of each gene on the first two principal correspondence

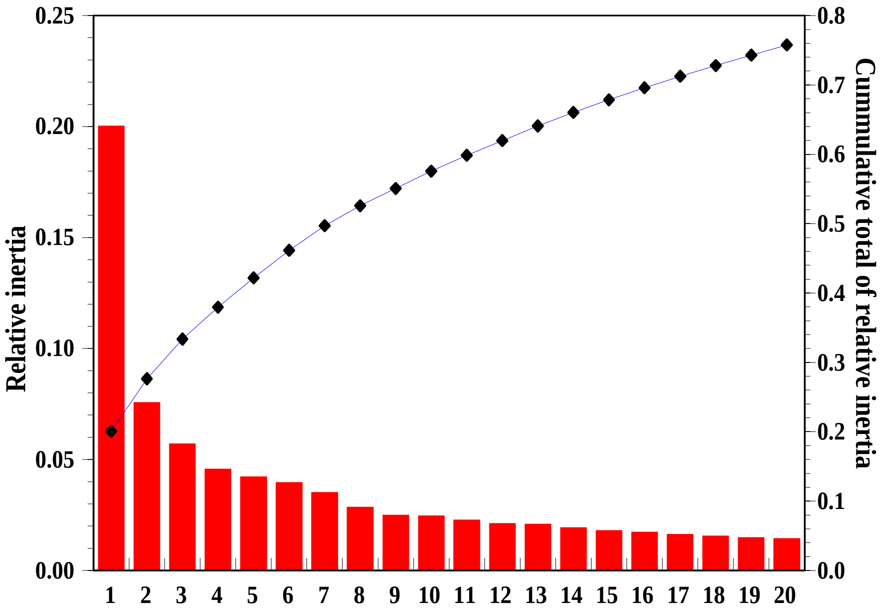

Figure 3-1. Each column represents a correspondence analysis factor from the analysis of the codon usage of 123 *L. lactis* genes ranked in decreasing order of the fraction of total variance or inertia in codon usage that it accounted for. The line represents the cumulative total of the inertia explained by the first 20 factors.

analysis axes is plotted in Figure 3-2. Genes with similar codon usage are plotted as neighbours. In Table 3-2 each gene is listed according to its position on axis 1. Many genes that have been known, or might be predicted to be highly expressed in other species or to preferentially use a subset of optimal codons (Andersson and Sharp 1996b; Gouy and Gautier 1982; Lloyd and Sharp 1993; Sharp *et al*. 1988) (e.g. genes encoding glycolytic enzymes and or ribosomal proteins) are located towards one extreme of axis 1. As the orientation of the axes is arbitrary, in this correspondence analysis, they are located on the left-hand side of the principal axis. Genes that one might predict to be moderately expressed such as genes involved in amino-acid biosynthesis lie around the centre of axis 1, while genes involved in regulation, lie at the other extreme of the principal axis.

The ordination of genes on the first four COA axes was examined for correlations with indices of codon usage and amino acid composition (e.g. EN_c, GC_{3s}, GC_{n3s}, GC, GRAVY, and Aromaticity). A summary of these correlations is presented in Table 3-3. There is a significant correlation between axis1 and both EN_c (r=0.746, p<0.0001), G+C (r=0.654, p<0.0001), G+C excluding synonymous third position i.e.GC_{n3s} (r=-0.803, p<0.0001), and aromaticity (r=0.447, p<0.001). The positive correlation with EN_c (see Figure 3-3) is caused by an increase in codon bias among the genes lying towards the left of axis 1, which includes genes coding for ribosomal proteins and glycolytic enzymes. While there is a correlation with G+C content, the lack of correlation with GC_{3s} (which might be expected to be influenced by a mutational bias) and the stronger correlation with GC_{n3s} (least likely to be influenced by a mutational bias) indicates that the CU variation represented by the first axis is not simply a consequence of G+C bias. Despite the lack of a correlation between GC_{3s} and axis 1, there is a significant correlation between axis 1 ordination and the frequency of C ending synonymous codons (r=0.39, p<0.001) which is counter-balanced by a negative correlation for G ending synonymous codons (r= -0.66, p<0.001). The correlation between the principal axis and G+C composition is partially due to a higher frequency of the Arg codon CGU and the increased frequency of the amino acids Ala (GCN) and Gly (GGN) among those genes (see Table 3-4). Together with the correlation with protein aromaticity this

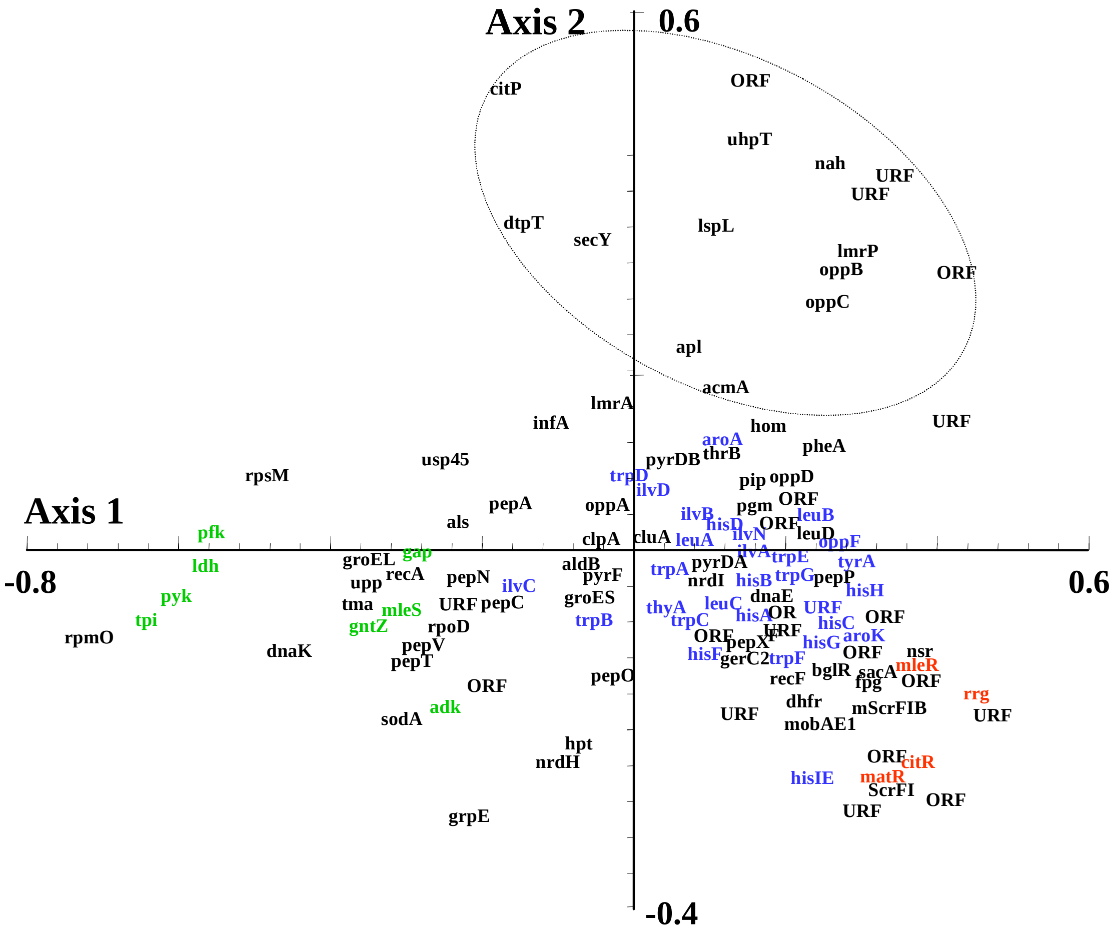

Figure 3-2 Correspondence analysis of codon usage variation among *L. lactis* genes. Each gene is plotted using its gene name at its coordinate on the first two axes produced by the analysis. The putatively highly expressed genes lie towards the left extreme of axis 1. The ellipse indicates a group of genes that putatively encode integral membrane proteins. Genes with encode regulatory genes are plotted in red, those genes encoding proteins that are involved in the metabolic pathways for amino acid production are plotted in blue, and those genes encoding glycolytic enzymes are plotted in green.

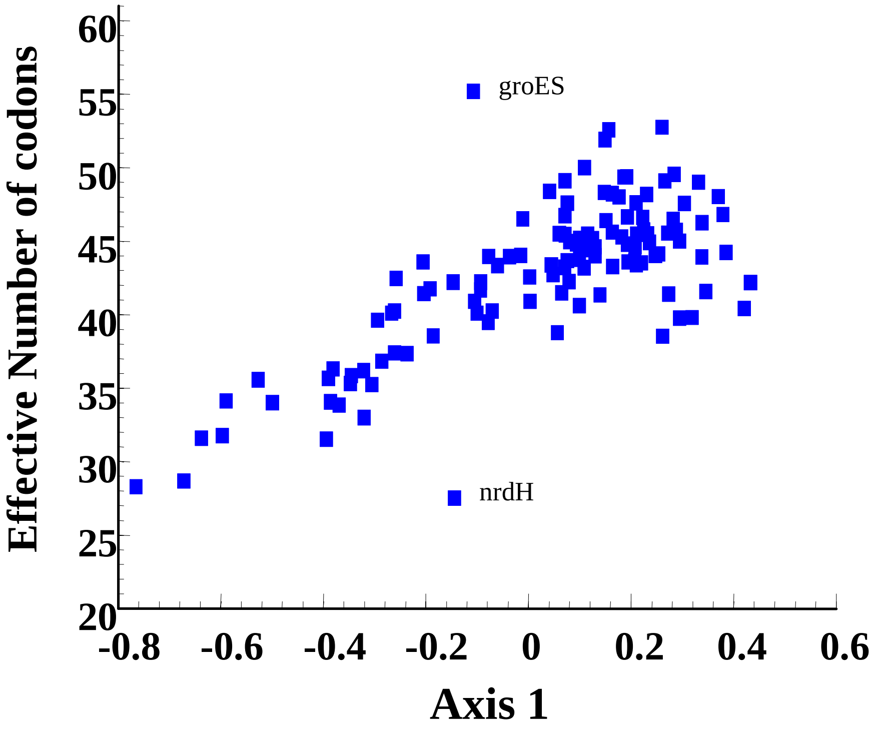

Figure 3-3. Correlation between the Effective Number of Codons for 123 *L. lactis* genes and correspondence analysis axis 1.

Table 3-2. *L. lactis* gene sequences**.**

Gene	GC3s	L	Gene description	ENc	Fop	Acc. #	PID	Reference:
_rpm_O	0.13	54	ribosomal protein L 15	28.3	0.84	X59250	g44072	JGM 137:2595
tpi	0.24	252	triose-phosphate isomerase	28.7	0.79	U07640	g537286	Micro 141:229
pyk	0.23	502	pyruvate 2-o-phosphotransferase	31.6	0.72	L07920	g308858	JBa 175: 2541
ldh	0.23	325	lactate dehydrogenase	31.8	0.71	M88490	g149424	JBa 174: 6956
pfk	0.24	340	D-fructose 6-phosphate 1-phosphotransferase	34.1	0.65	L07920	g308857	JBa 175: 2541
_rps_M	0.14	51	ribosomal protein S13	35.6	0.61	X59250	g44077	JGM 137:2595
_dna_K	0.20	607	heat shock protein	34.0	0.68	X75428	g450684	Gene 142:,91
upp	0.21	211	uracil phosphoribosyl transferase	31.5	0.64	X73329	g599847	JBa 176:6475
gntZ	0.24	472	6-phosphogluconate dehydrogenase	35.7	0.61	U74322	g1857246	Unpublished
_gro_EL	0.14	542	heat shock protein *cpn60*	34.1	0.59	X71132	g287871	Gene 127: 121
tma	0.21	695	transmembrane ATPase	36.3	0.60	X69123	g44027	CurrG 25: 379
mleS	0.15	521	malolactic enzyme, decarboxylation of L-malate	33.9	0.61	X71897	g467569	FEMS 116:79
_sod_A	0.20	206	superoxide dismutase	35.3	0.57	U17388	g755210	JBa 177: 5254
_rec_A	0.21	365	SOS induction gene	35.8	0.60	M88106	g551877	AEM 58: 2674
_pep_T	0.20	413	tripeptidase	36.2	0.58	L27596	g495046	JBa 176:2854
pepV	0.21	472	dipeptidase	33.0	0.58	U78036	g2160707	JBa 179: 3410
gap	0.20	337	glyceraldehyde-3-phosphate dehydrogenase	37.1	0.55	L36907	g806486	Micro 141:1027
adk	0.18	215	adenylate kinase	35.2	0.58	X59250	g44074	JGM 137:2595
_usp_45	0.18	461	secreted protein	39.6	0.56	A17083	g512521	Gene 95: 155
_rpo_D	0.23	340	RNA polymerase sigma factor	36.8	0.56	X71493	g403013	BBB 57: 88
URF	0.30	849	unknown	40.1	0.56	D38040	d1007812	Unpublished
_pep_N	0.23	846	amino peptidase	37.4	0.57	M87840	g149464	NMDJ 47:54
als	0.24	554	alpha-acetolactate synthase	40.3	0.53	L16975	g473902	AEM: 60 1390
_grp_E	0.22	179	heat shock protein	42.5	0.63	X76642	g435491	JGM 139:3253
ORF	0.29	211	homologous to *B. stearthermophilus fms*	40.8	0.41	L36907	g806487	Micro 141:1027
_pep_C	0.24	436	cysteine aminopeptidase	37.4	0.60	M86245	g149364	AEM 59 :330
_pep_A	0.26	355	glutamyl-aminopeptidase	43.6	0.50	X81089	g1072381	Micro 141:2873
_cit_P	0.38	442	citrate permease	41.5	0.50	M58694	g149369	JBa 172: 5789
_ilv_C	0.24	344	acetohydroxy acid isomeroreductase	41.8	0.53	M90761	g149434	JBa 175: 4383
_dtp_T	0.25	463	tri-peptide transporter	38.6	0.48	U05215	g451072	JBC 269:11391
_inf_A	0.39	72	initiation factor IF-1	42.2	0.49	X59250	g581299	JGM 137:2595
_nrd_H	0.21	72	glutaredoxin-like protein	27.5	0.51	X92690	e209004	JBC 271: 8779
hpt	0.14	183	hypoxanthine guanine phosphoribosyltransferase	40.9	0.53	X67015	g49105	CurrG 25: 379
_gro_ES	0.23	94	heat shock protein cpn10	55.2	0.48	X71132	g287870	Gene 127: 121
_ald_B	0.20	236	alpha-acetolactate decarboxylase	40.1	0.50	X82620	e124050	AEM 62: 2641
_clp_A	0.16	137	ATP-dependent protease	37.0	0.48	L36907	g806484	Micro 141:1027
_trp_B	0.22	402	tryptophan synthase beta subunit	42.2	0.47	M87483	g149521	JBa 174: 6563
_sec_Y	0.26	439	export protein	41.7	0.50	X59250	g44073	JGM 137:2595
_opp_A	0.22	600	oligopeptide binding protein	39.5	0.49	L18760	g308854	JBa 175: 7523
_pyr_F	0.22	237	omp decarboxylase	44.0	0.52	X74207	e264705	JBa 176: 3975
_pep_O	0.18	627	endopeptidase	40.2	0.49	L18760	g308855	Unpublished
_lmr_A	0.25	590	multidrug resistance protein	43.4	0.49	U63741	g1890649	PNAS 93:10668
_trp_D	0.26	335	phosphoribosyl anthranilate transferase	44.0	0.48	M87483	g149518	JBa 174: 6563
_clu_A	0.24	1243	sex factor aggregation protein	44.1	0.44	U04468	g458234	MM 12: 655
_ilv_D	0.26	570	dihydroxy acid dehydrase	46.5	0.45	M90761	g149431	JBa 174: 6580
_pyr_DB	0.28	311	dihydroorotate dehydrogenase B	40.9	0.44	X74207	e264499	JBa 176: 3975
_thy_A	0.24	279	thymidylate synthase	42.6	0.47	M33770	g149513	AEM 56: 2156
_trp_A	0.21	253	tryptophan synthase alpha subunit	43.4	0.47	M87483	g149522	JBa 174: 6563
apl	0.29	242	alkaline phosphatase	48.4	0.44	Z29065	g435296	Unpublished
_ilv_B	0.22	575	acetolactate synthase	42.8	0.43	M90761	g149432	JBa 174: 6580
_nrd_I	0.24	140	homologous to ORF2 in *E. coli nrd*EF operon	38.8	0.48	X92690	e209005	JBC 271: 8779
_pyr_DA	0.21	311	dihydroorotate dehydrogenase A	45.5	0.44	X74206	g511015	JBa 176: 3975

Table 3.2 *L. lactis* gene sequences(cont.)

Gene	GC3s	L	Gene description	EN_c	F_{op}	Acc. #	Pid	Reference:
ORF	0.16	211	homologous to *E. coli vac*B gene	41.5	0.55	M90760	g149373	JBa 174: 6571
***leu*C**	0.21	460	alpha isopropylamate dehydratase	45.4	0.41	M90761	g149428	JBa 174: 6580
***thr*B**	0.27	296	threonine biosynthesis	46.8	0.42	X96988	e234079	JBa 178: 3689
***trp*C**	0.27	264	indoleglycerol phosphate synthase	49.1	0.49	M87483	g149519	JBa 174: 6563
***aro*A**	0.23	430	5-enolpyruvylshikimate-3-phosphate synthase	43.7	0.42	X78413	g683583	MGG 246:119
***lsp*L**	0.23	143	signal peptidase type II	43.2	0.47	U63724	g1480916	Unpublished
***his*F**	0.21	244	cyclase	42.3	0.45	M90760	g149383	JBa 174: 6571
***acm*A**	0.23	437	N-acetylmuramidase	47.6	0.43	U17696	g755216	JBa 177:1554
***his*D**	0.23	431	histidinol dehdrogenase	45.0	0.43	M90760	g149377	JBa 174: 6571
***leu*A**	0.19	513	alpha-isopropylmate synthase	44.8	0.42	M90761	g149426	JBa 174: 6580
***trp*G**	0.24	198	anthranilate synthase beta subunit	45.2	0.41	M87483	g551880	JBa 174: 6563
URF	0.19	349	ORF located near C-terminal of *als*	40.6	0.47	L16975	g473901	AEM 60: 1390
***ger*C2**	0.18	252	analine simulated germination in *B. subtilis*	43.8	0.45	L14679	g410740	JBa 174: 3577
***pep*X**	0.24	763	X-prolyl dipeptidyl aminopeptidase	43.2	0.43	M35865	g149467	AEM 57: 45
***ilv*N**	0.21	158	acetolactate synthase	45.5	0.36	M90761	g149433	JBa 174: 6580
***uhp*T**	0.35	215	hexose phosphate transport	50.0	0.41	X71493	g403014	BBA 1216:115
ORF	0.29	119	homologous of *E. coli yce*E gene	44.6	0.44	X62621	g44067	Gene 119: 145
pgm	0.21	221	beta-phosphoglucomutase	44.5	0.39	Z70730	e236074	Micro 143: 855
pip	0.24	901	required for phage infection	45.2	0.43	L14679	g308861	JBa 174: 3577
***ilv*A**	0.22	441	threonine synthase	44.0	0.41	M90761	g149435	JBa 174: 6580
hom	0.22	428	threonine biosynthesis	44.6	0.41	X96988	e234078	JBa 178: 3689
***dna*E**	0.20	393	DNA primase	41.4	0.40	D14690	d1004027	BBB 59:73
ORF	0.29	119	homologous of *E. coli yce*E gene	44.6	0.44	X62621	g44067	Gene 119: 145
pgm	0.21	221	beta-phosphoglucomutase	44.5	0.39	Z70730	e236074	Micro 143: 855
pip	0.24	901	required for phage infection	45.2	0.43	L14679	g308861	JBa 174: 3577
***ilv*A**	0.22	441	threonine synthase	44.0	0.41	M90761	g149435	JBa 174: 6580
hom	0.22	428	threonine biosynthesis	44.6	0.41	X96988	e234078	JBa 178: 3689
***dna*E**	0.20	393	DNA primase	41.4	0.40	D14690	d1004027	BBB 59:73
ORF	0.26	156	homologous to *B. subtilis ywbgh*	51.9	0.39	L36907	g806485	Micro 141:1027
***his*B**	0.27	200	dehydratase	51.9	0.39	M90760	g149379	JBa 174: 6571
***his*A**	0.23	239	isomerase	48.3	0.41	M90760	g149382	JBa 174: 6571
ORF	0.25	347	homologous ORF in *B. subtilis* also 5' to *grp*E	46.4	0.42	X76642	g435490	JGM 139:3253
URF	0.35	263	ORF located at N-terminal of *his*D gene	52.6	0.43	M90760	g149378	JBa 174: 6571
***trp*E**	0.27	456	anthranilate synthase alpha subunit	48.2	0.44	M87483	g149516	JBa 174: 6563
***opp*D**	0.22	338	ATP binding protein	43.3	0.44	L18760	g308850	JBa 175: 7523
***trp*F**	0.22	351	phosphoribosyl anthranilate isomerse	45.6	0.43	M87483	g149520	JBa 174: 6563
ORF	0.42	299	proteinase activation protein	48.0	0.38	M26694	g623055	JBa 171: 2789
***leu*D**	0.24	191	alpha isopropylamate dehyratase	45.3	0.43	M90761	g149429	JBa 174: 6580
dhfr	0.24	168	dihydrofolate reductase	49.4	0.49	X60681	g530024	AEM 61: 561
***mob*AE1**	0.28	584	mobilisation protein	49.4	0.43	X89779	e191395	JBC 270:26092
***leu*B**	0.21	345	3-iso-propylmalate dehydrogenase	44.8	0.38	M90761	g912423	JBa 174: 6580
***hisI*E**	0.18	212	hydrolase pyrophosphohydrolase	46.7	0.44	M90760	g149384	JBa 174: 6571
***rec*F**	0.21	357	DNA repair	43.6	0.43	X89367	e187696	Gene 170: 151
***opp*F**	0.22	319	ATP binding protein	43.7	0.39	L18760	g308851	JBa 175: 7523
***his*G**	0.17	208	phosphoribosyl-ATP synthetase	43.4	0.42	M90760	g149376	JBa 174: 6571
***phe*A**	0.28	279	prephenate dehydratase	44.6	0.35	X78413	g683585	MGG 246:119
URF	0.25	195	located at the C-terminal of *pep*C	47.6	0.42	M86245	g293011	AEM 59 :330
***opp*C**	0.26	294	transmembrane protein	45.5	0.37	L18760	g308853	JBa 175: 7523
***bgl*R**	0.20	265	regulator, involved in beta-glucoside utilisation	43.5	0.45	L27422	g551875	JBa 176: 5681
***pep*P**	0.25	352	aminopeptidase P	46.6	0.39	Y08842	e276466	Unpublished
Nah	0.29	379	sodium-hydrogen antiporter	45.8	0.35	U78036	g599848	JBa 176: 6457
***tyr*A**	0.21	354	prephenate dehydrogenase	45.5	0.38	X78413	g683582	MGG 246:119
***opp*B**	0.27	319	transmembrane protein	48.2	0.34	L18760	g308852	JBa 175: 7523

Table 3-9. *L. lactis* gene sequences **(cont.).**

Gene	GC3s	L	Gene description	EN_c	F_{op}	Acc. #	Pid	Reference:

***his*C**	0.21	360	aminotransferase	44.9	0.39	M90760	g149374	JBa 174: 6571
ORF	0.23	259	homologous to *E. coli xyl*G gene	44.0	0.41	M90761	g149430	JBa 174: 6580
***lmr*P**	0.21	408	*lmr*P integral membrane protein	44.2	0.33	X89779	g1052754	JBC 270: 26092
URF	0.28	155	5’ to a *L. lactis* 16s RNA	52.8	0.41	X65713	g433942	JGM 139:2009
***aro*K**	0.20	162	shikimate kinase	38.5	0.38	X78413	g683584	MGG 246:119
***his*H**	0.24	202	amidotransferase	49.1	0.37	M90760	g149381	JBa 174: 6571
M*Scr*FIB	0.20	360	5-methylcytosine methyltransferase	41.4	0.35	L12227	g149495	Gene 136: 205
URF	0.26	119	hypothetical protein 5’ to *trp*E	45.6	0.40	M87483	g551879	JBa 174: 6563
***sac*A**	0.19	473	sucrose-6-phosphate hydrolase	46.5	0.40	M96669	g149490	Gene 121: 55
***mat*R**	0.27	599	maturase-related protein	49.6	0.42	X89922	e191396	MM. 21, 45-53
ORF	0.23	305	homologous to *E. coli* DNA primase	45.8	0.41	D10168	g216732	BBB 57: 88
***scr*FI**	0.13	389	*scr*FI methylase	39.8	0.39	M87289	g149493	AEM 59: 777
fpg	0.23	273	formamidopyrimidine-DNA glycosylase	45.0	0.45	X74298	g433584	MUK 141: 411
URF	0.36	120	located at the C-terminal of *apl*	47.6	0.31	Z29065	g435297	Unpublished
ORF	0.21	91	homologous to a *B. subtilis* hypothetical protein	39.8	0.38	M90762	g535664	JBa 171: 3108
***mle*R**	0.23	291	positive regulator	49.0	0.41	M90762	No PID	JBa 171:3108
***cit*R**	0.28	112	citrate permease P	46.3	0.41	S77101	g913983	MGG 246:590
ORF	0.21	328	homologous to histidyl-tRNA synthetase	43.9	0.40	M90760	g149375	JBa 174: 6571
nsr	0.17	318	nisin resistance protein	41.6	0.35	M37002	g149456	AEM 57: 804
ORF	0.19	262	homologous to aminoglycoside phosphotransferase	48.1	0.38	M90760	g149380	JBa 174: 6571
URF	0.22	614	ORF located near N-terminal of *dna*K	46.8	0.33	X76642	g435493	JGM 139:3253
ORF	0.26	198	homologous to *E. coli* cell division gene *fts*W	44.3	0.33	X62621	g44069	Gene 119: 145
rrg	0.19	99	regulator of glucosyltransferase expression	40.4	0.36	L14679	g410739	JBa 174: 3577
URF	0.16	269	ORF located near C-terminal of *his*IE	42.2	0.33	M90760	g149385	JBa 174: 6571

Genes are listed in the order of their position on correspondence analysis axis 1. Gene is the gene name of each sequence, gene names in blue are partial sequence. Where a gene has not been identified but has homology with another sequence it is labelled as an ORF, where it was not possible to identify a homologous sequence it is labelled as a unidentified reading frame URF. L is the length of the gene in codons. GC_{3s} is the G+C content at silent positions. EN_c is the effective number of codons used in a gene. F_{op} is the frequency of optimal codons. Acc. # is the GenBank/EMBL/DDBJ accession number. PID is the GenBank/EMBL/DDBJ protein identification number. References are given in an abbreviated notation, JBa, J. Bact.; NMDJ, Netherlands Milk and Dairy Journal; BBB, Biosci. Biotech. Biochem.; CurrG, Current Genetics; MM, Mol. Micro.; BBA, Biochim. Biophys. Acta.; JBC: J. Biol. Chem.; JGM: J. Gen. Microbiol.; MUK, Microbiology-UK; MGG, Mol. Gen. Genet.; EJB, Euro. Bio. J.; AEM, App. Env. Micro.; Micro, Microbiology; FEBS, FEBS Letters; FEMS, FEMS. Letters.

	EN_c	GC_{3s}	GC_{n3s}	GC	Gravy	Aromo	CBI	Fop
Axis 1	0.746	0.139	-0.803	-0.654	-0.026	0.447	-0.956	-0.943
Axis 2	0.151	0.479	0.071	0.315	0.868	0.211	-0.099	-0.172
Axis 3	-0.040	-0.042	0.004	-0.026	-0.239	0.019	0.049	0.075
Axis 4	-0.227	-0.577	-0.164	-0.156	-0.106	-0.356	0.077	0.028

Table 3-3. Correlation between the codon usage and amino acid usage indices, ENc GC3s, GCn3s GC GRAVY, Aromaticity, CBI and Fop (see Chapter 2) and the first four correspondence analysis axes from an analysis of 123 *L. lactis gene*s.

Table 3-4. Putative optimal and rare codons of *L. lactis*

		High		Low				High		Low	
		RSCU	N	RSCU	N			RSCU	N	RSCU	N
Phe	UUU	0.89	(24)	1.81	(123)	Ser	UCU	1.57	(22)	1.42	(32)
	UUC	1.11	(30)	0.19	(13)		UCC	0.00	(0)	0.36	(8)
Leu	UUA	0.17	(3)	2.37	(81)		UCA	3.57	(50)	2.04	(46)
	UUG	2.20	(40)	1.02	(35)		UCG	0.07	(1)	0.36	(8)
	CUU	3.19	(58)	1.05	(36)	Pro	CCU	1.33	(15)	1.76	(26)
	CUC	0.44	(8)	0.35	(12)		CCC	0.00	(0)	0.75	(11)
	CUA	0.00	(0)	0.82	(28)		CCA	2.49	(28)	1.02	(15)
	CUG	0.00	(0)	0.38	(13)		CCG	0.18	(2)	0.47	(7)
Ile	AUU	1.16	(43)	1.68	(93)	Thr	ACU	2.52	(53)	1.95	(40)
	AUC	1.84	(68)	0.38	(21)		ACC	0.00	(0)	0.15	(3)
	AUA	0.00	(0)	0.94	(52)		ACA	1.43	(30)	1.27	(26)
Met	AUG	1.00	(34)	1.00	(33)		ACG	0.05	(1)	0.63	(13)
Val	GUU	2.69	(92)	1.80	(32)	Ala	GCU	2.49	(112)	1.44	(36)
	GUC	0.20	(7)	0.68	(12)		GCC	0.24	(11)	0.52	(13)
	GUA	1.08	(37)	1.01	(18)		GCA	1.18	(53)	1.64	(41)
	GUG	0.03	(1)	0.51	(9)		GCG	0.09	(4)	0.40	(10)
Tyr	UAU	0.30	(5)	1.81	(75)	Cys	UGU	1.67	(5)	1.71	(12)
	UAC	1.70	(28)	0.19	(8)		UGC	0.33	(1)	0.29	(2)
TER	UAA	3.00	(6)	3.00	(4)	TER	UGA	0.00	(0)	0.00	(0)
	UAG	0.00	(0)	0.00	(0)	Trp	UGG	1.00	(8)	1.00	(16)
His	CAU	0.61	(7)	1.78	(24)	Arg	CGU	5.32	(55)	1.00	(9)
	CAC	1.39	(16)	0.22	(3)		CGC	0.68	(7)	0.44	(4)
Gln	CAA	2.00	(34)	1.72	(43)		CGA	0.00	(0)	1.11	(10)
	CAG	0.00	(0)	0.28	(7)		CGG	0.00	(0)	0.44	(4)
Asn	AAU	0.71	(24)	1.79	(92)	Ser	AGU	0.14	(2)	1.29	(29)
	AAC	1.29	(44)	0.21	(11)		AGC	0.64	(9)	0.53	(12)
Lys	AAA	1.90	(93)	1.63	(120)	Arg	AGA	0.00	(0)	2.56	(23)
	AAG	0.10	(5)	0.37	(27)		AGG	0.00	(0)	0.44	(4)
Asp	GAU	1.17	(49)	1.52	(61)	Gly	GGU	2.54	(94)	1.14	(30)
	GAC	0.83	(35)	0.47	(19)		GGC	0.30	(11)	0.61	(16)
Glu	GAA	1.95	(119)	1.62	(76)		GGA	1.03	(38)	1.45	(38)
	GAG	0.05	(3)	0.38	(18)		GGG	0.14	(5)	0.80	(21)

The high bias and low bias sets of genes contain 1529 and 1764 codons respectively. Those codons that occur significantly more often ($p<0.01$) in the highly biased dataset relative to the lower biased dataset are putatively considered optimal, and are indicated in red. The codon CAA (Gln) is significant at $p<0.05$, and is marked in purple. Those codons that are rare or absent as defined by Sharp *et al.* 1990, i.e. RSCU < 0.10, are indicated by blue.

that amino acid composition varies systematically along axis 1.To exclude the possibility that the variation in CU identified by axis 1 was influenced by variation in amino acid composition a correspondence analysis of RSCU (i.e. codon usage with amino acid composition normalised) of the same dataset was calculated (data not shown). The positions of the genes on the principal axis in both correspondence analyses was strongly conserved ($r=0.965$, $p<0.001$) and the correlation between the principal axis of the RSCU correspondence analysis and aromaticity remained significant though slightly lower ($r=0.305$, $p<0.01$). A correlation between amino acid composition and gene expressivity in *E. coli* has been previously reported (Lobry and Gautier 1994) and a similar but weaker correlation has already been presented for *S. typhimurium* in this thesis.

The second COA axis was significantly correlated ($p<0.001$) with the GRAVY score, an index of the hydrophobicity of the conceptually translated gene products. The second axis differentiates a group of integral-membrane proteins, delineated by an ellipse in Figure 3-2. This group includes ORF (PID:g44067) - homologous to *E. coli* transmembrane protein *yce*E, *cit*P - citrate permease, *uhp*T – hexose phosphate transferase, *nah* – sodium ion transporter, *dtp*P – a tripeptide transporter, *lsp*L – signal peptidase, *opp*C and *opp*B – oligopeptide transport transmembrane proteins, *sec*Y – integral membrane involved in protein export, acmA – cell wall protein, ORF (PID:g44069) – homologous to *E. coli fts*W a cell division and integral membrane protein, and *apl* – a periplasmic cell division protein. This group of genes also includes the 119 codon URF (PID:g551879) a hypothetical protein from the *trp* operon, and the 120 codon URF (PID:g435297) located at the C-terminal of *apl*. While no homologues could be identified for these ORFs their usage of codons is consistent with their being integral membrane proteins. The correlation between correspondence analysis axis 2 and GRAVY is not observed if RSCU data is used in the COA.

A more detailed examination of the relationship between EN_c and axis 1 (see Figure 3-3) reveals two outliers, *gro*ES and *hrd*H. The calculations of EN_c for short sequences (i.e. < 100 codons) or genes with unusual amino acid composition can result in large standard errors (Wright 1990). Both these genes are short and compared with other genes from the same dataset they had biased amino acid compositions. *Hrd*H (72 AA) encodes a protein with the lowest proportion of 6-fold amino acids and *gro*ES (92 AA) has the third lowest proportion of

2-fold synonymous amino acids. If these genes are excluded the correlation coefficient would increase to r=0.80.

3.4.2.3 Optimal and non-optimal codons

Optimal codons have been defined as those codons which occur significantly more frequently in highly expressed genes than in genes with lower levels of expression (Lloyd and Sharp 1991; Sharp and Devine 1989). To identify the putative optimal codons, codon usage is contrasted between two groups of genes taken from the extremes of axis 1. Both groups contained 6 sequences (5% of the dataset), the significance of the difference in codon usage between the two groups was evaluated using a 2x2 Chi squared contingency test. This allowed the identification of 17 optimal codons for 16 of the 18 synonymously variable amino acids. No optimal codon was identified for Thr, or the relatively rare amino acid Cys, while two optimal codons (UUG and CUU) were identified for Leu (see Table 3-4). In the group of more highly biased genes, CAA was used exclusively to encode the 34 Gln amino acids, although the frequency of this codon was only significant at $p<0.05$ it was tentatively assigned as optimal. The putative optimal codons included the universally optimal codons UUC, UAC, AUC, GAC, AAC, and GGU (Sharp and Devine 1989) and the codons CUU, GUU, CAA, AAA, GAA, CGU, GCU, and CCA which have been identified as optimal codons in *B. subtilis* (see Table 4-6).

In those species where selection for optimal translation is choosing systematically between codons, some codons are infrequently used in highly expressed genes. These “rare” codons have been defined as codons that are utilised at less than 10% of their expected frequency if random codon usage was assumed – i.e. RSCU < 0.1 (Sharp *et al*. 1990). In *L. lactis* 17 codons were classified as rare according to this definition (see Table 3-4). The majority of these codons such as CCC, UCG, and UCC are G+C rich, but the rare codons also included the more A+T rich codons CUA and AUA, which have already been noted as being infrequent in the dataset. Other codons classified as rare included AUA, CCC, CGA, CGG, and AGG, these codons are also classified as rare in *E. coli* and *B. subtilis* (see Table 4-6).

The distribution of the gene types on axis 1 and the correlation between axis 1 and EN_c, are strong indicators that the major trend in the codon usage variation represented by this axis, is a bias towards a set of preferred codons. The overlap between the putative optimal and putative rare codons of *L. lactis* with *B. subtilis* supports the hypothesis that these preferred codons are being selected for some type of translation efficiency. The codons identified as being used significantly more often in Table 3-4 overlap with those identified as being optimal in other species (see Table 4-6). Therefore, the codons found significantly more often will be considered to be optimal codons. Genes with a higher proportion of optimal codons will be tentatively considered to have an increased expression level.

Correspondence analysis treats the rows (genes) and columns (codons) in a symmetric fashion (Lebart, Morineau and Warwick 1984) and it allows the rows and columns to be represented in the same lower dimensional space. Thus far only genes have been projected onto the two principal axes, but the projection of codons can be equally informative. The distance each codon is from the origin is a relative indication of the influence of each codon on the principal trends in the variation. In Figure 3-4 the aforementioned optimal codons and rare codons are clustered in two groups on either side of the principal axis. The optimal codon with the largest influence on the principal axis is the Arg codon CGU, while the rare codon with the greatest influence is the Ile codon AUA. Interestingly, although the codons AAA and CAA are classified as optimal neither has a strong influence on the principal trend in the variation. A closer examination of Table 3-4 reveals that while occurring statistically more often among the high bias genes, these codons are also the preferred codons in the low bias genes.

F_{op} represents the degree to which a gene has adapted its codon usage towards the subset of these optimal codons. F_{op} was calculated for each gene in the dataset using the 17 optimal codons described above and is presented in Table 3-4. As optimal codons were only identified for 16 amino acids (no optimal codons were identified for Cys and Thr), the frequency was calculated only for these 16 amino acids. The index is tabulated in Table 3-2. As optimal codons were identified from the codon usage of two groups of genes from the extremes of axis 1, it is unsurprising that F_{op} is correlated with axis 1 ($r=-0.943$, $p<0.001$). However, the strength of the correlation confirms that the variation in the usage of 17 (optimal) codons can emulate the most important correspondence analysis

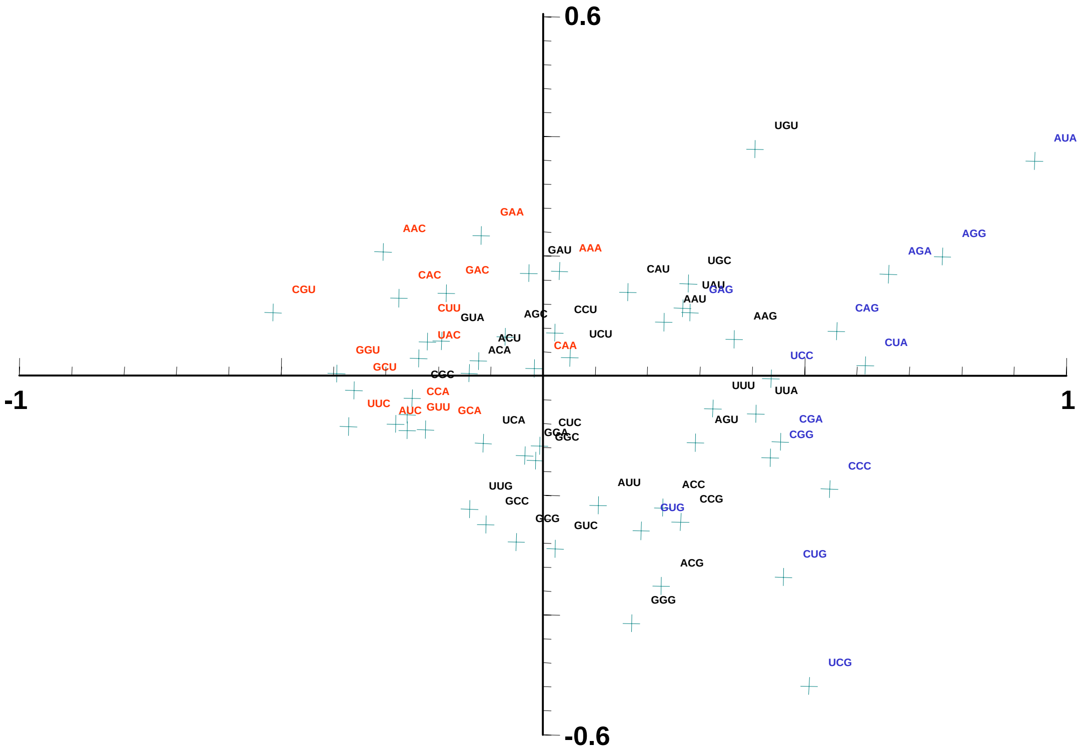

Figure 3-4. Factor map of the 59 *L. lactis* synonymous codons on the two principal axes produced by the analysis. Those codons which were identified as optimal are indicated by red, and rare codons are indicated by blue.

vector, representing 20% of the variation in total usage in this dataset. The modified F_{op} index, which factors in the frequency of rare codons, had a higher correlation with axis 1 (r=0.957, p<0.001) as did the codon bias index (r=0.956, p<0.001).

3.4.2.4 Inferred Levels of gene expression

Codon usage indices have been used to infer relative expression levels under the hypothesis that genes with a high proportion of optimal codons are highly expressed. The gene encoding the ribosomal protein L 15 (*rpm*O) had the highest F_{op} (0.84) and greatest influence on the first axis. The expression level of *rpm*O in *L. lactis* has not been determined but its *E. coli* homologue has been classified as being very highly expressed (Sharp and Li 1986) and its *B. subtilis* homologue *rmlO* has a high F_{op} (0.67). Three genes sequences, *adk* (F_{op}=0.58), *inf*A (F_{op}=0.49) and *rps*M (F_{op}=0.61) are co-transcribed from the same operon as *rpmO*. Koivula and Hemila (1991) speculated that based on calculated free energy for ribosomal binding sites (RBS) *rpsM* would have higher expression than *adk*, this supports their relative F_{op} values. Koivula and Hemila also predicted that *infA* would have higher expression than *adk* because of its stronger RBS, however *infA* has a lower F_{op}. The translation initiation codon for *infA* is UUG, which is known to be less efficient than AUG and this may negate some of the influence of a strong RBS (Koivula and Hemila 1991). Triosephosphate isomerase (TPI), which has the second highest F_{op} (0.79), catalyses the interconversion of dihydroxyacetone phosphate and glyceraldehyde 3-phosphate, which is the first common step for the catabolism of monosaccharide moieties of lactose via the EMP pathway (Cancilla *et al*. 1995a). TPI has been previously described as one of the most abundant enzymes in the bacterial cell, comprising approximately 2% of soluble protein (Noltmann 1972).

The genes *ldh* (F_{op}=0.71), *pyk* (F_{op}=0.72), and *pfk* (F_{op}=0.65) are co-transcribed from the tricistronic *las* operon. Their products catalyse steps of the EMP pathway and have high F_{op} values. They have been predicted to be highly expressed (Llanos *et al*. 1993; Llanos, Hillier and Davidson 1992). Using enzyme activity assays the concentrations of their gene products have been estimated. Pyruvate kinase (PYK) -300 pMol/mg and lactate dehydrogenase (LDH) -280 pMol/mg were found to be more abundant than phosphofructokinase PFK - 75 pMol/mg (Llanos *et al*. 1993). These concentrations support their high F_{op} values, with *pyk* and *ldh* having similar F_{op} values, both slightly higher than *pfk*. Whether or not the difference of 0.06 in the F_{op} between *pyk* and *pfk* reflects a 4:1 difference in expression levels is less certain however.

Glyceraldehyde-3-phosphate dehydrogenase catalyses the reversible oxidative phosphorylation of glyceraldehyde 3-phosphate to produce 1,3-biphosphoglycerate, using NAD+ as a coenzyme. This enzyme has an important role in glycolysis, forming a central link between ATP-production and NAD+ reduction. A decrease in the glycolytic activity of starved *L. lactis* cells has been shown to be the result of a diminished level of glyceraldehyde-3-phosphate dehydrogenase (Poolman 1993). It has been suggested that the *L. lactis gap* gene is strongly expressed (Cancilla, Hillier and Davidson 1995b), however this suggestion was based on a comparison of the codon usage of *gap, ldh, tpi, pfk,* and *pyk*, with the codon usage of other chromosomal genes. The F_{op} for *gap* (F_{op}=0.55) is not particularly high. While these five glycolytic genes do have a markedly different codon usage from chromosomal genes as a whole it is obviously overly simplistic to compare codon usage by automatically grouping these genes by function. The difference in codon usage is primarily due to *ldh, tpi, pfk,* and *pyk,* while *gap* displays a codon usage more reminiscent of a moderately expressed gene.

Heat shock proteins (HSPs) have highly conserved amino acid sequences. In addition to being induced by stress, they are also molecular chaperones, mediating the folding and assembly of cellular proteins (Gething and Sambrook 1992; Langer *et al*. 1992). The HSPs *dnaK* (F_{op}=0.68), *groEL* (F_{op}=0.59)*, grpE* (F_{op}=0.63)*,* and *groES* (F_{op}=0.48) have been cloned from *L. lactis*. The HSP *dnaK*, which has the highest F_{op}, is essential for cell viability and DNA replication and is among the most abundant HSPs in the cell (Barril *et al*. 1994). In *L. lactis,* heat shock induces *dnaK* up to three-fold (Eaton *et al*. 1993). *DnaK* is transcribed as a monocistronic transcript with *grpE* and an ORF (PID:g435490). The ORF has a homologue in the *B. subtilis grp*E-ORF-*dna*K operon (Eaton, Shearman and Gasson 1993). *GrpE* has a high F_{op} (0.63) while the ORF has a much lower F_{op} (0.42). As in *E. coli* and *B. subtilis,* the *L. lactis gro*EL has a higher F_{op} than *groES*. A fourth URF (PID:g435493) is located immediately downstream of the *dnaK*-ORF-*grpE* operon at a position occupied by *dnaJ* in *E. coli* and *B. subtilis,* this URF has a remarkably low F_{op} (0.31). Despite this, and the absence of significant sequence similarity with any known ORF, the length of this ORF (614 codons) strongly supports its identification as a transcribed ORF.

It has been shown that lactococci are able to utilise different exogenous pyrimidine sources, including uracil. The key step in the salvage of uracil is the reaction of uracil and 5-phosphoribosyl-α-1-pyrophosphate, a reaction that is catalysed by uracil-phospho-ribosyltransferase encoded by *upp* (F_{op}=0.64). The strong F_{op} of *upp* is supported by the presence of a particularly stable transcription terminator and a strong promoter, which when cloned into *E. coli* can over-express *upp* (Martinussen and Hammer 1994). Auxotrophic complementation experiments demonstrated that uracil is the sole pyrimidine source only in the presence of the *upp* product.

The gene *rpm*G encodes the ribosomal protein L33, but was removed from original dataset because it was shorter than 50 codons. Unlike the other *L. lactis* ribosomal genes *rpm*O and *rps*M, *rpm*G has a surprisingly low F_{op} (0.39) which ranks it amongst the lowest 20% of F_{op} values in the dataset. Interestingly, the *rpm*G homologue in *B. subtilis* also has a low F_{op}. Apparently, the selective pressures which increase the frequency of optimal codons in the *E. coli rpm*G gene are unable to withstand genetic drift present in these particular Gram-positives. Whether this is because its product has a lower expression level is less certain. In *B. subtilis*, the *rpm*G gene may be subjected to some other type of selective pressure, an investigation of the divergence of gene sequences between *B. subtilis* and *Bacillus licheniformis* found that *rpm*G had a markedly lower synonymous substitution rate (Sharp, Nolan and Devine 1995a). The reason for this low rate is unclear, it could have been due to the number of sites sampled or the general genome location of the *B. subtilis rpm*G gene.

3.4.2.5 Correlations with the codon usage of homologous B. subtilis genes

All 124 *L. lactis* ORFs were used to search for *B. subtilis* homologues using the methodology described in Methods. 71 homologous genes were identified (Table 3-5), and these exhibit a range of codon usage patterns in *L. lactis* (Figure 3-4). This was used to infer that the 123-gene *L. lactis* dataset used to identify optimal codons contained sequences that were representative of genes known to be expressed at a range of levels in other species. The ENc calculated from

Table 3-5 Codon usage of *L. lactis* and *B. subtilis* homologues

LL	L	GC_{3s}	F_{op}	CAI	BS	L	GC_{3s}	F_{op}	CAI
adk	215	0.18	0.58	0.60	*adk*	217	0.33	0.40	0.51
als	554	0.24	0.53	0.50	*als*S	540	0.41	0.43	0.46
*aro*A	430	0.23	0.42	0.34	*aro*A	358	0.43	0.49	0.49
citP	442	0.38	0.50	0.39	*yuf*R	448	0.39	0.35	0.42
dhfr	168	0.24	0.49	0.35	*dfra*	168	0.42	0.35	0.40
*dna*E	393	0.20	0.40	0.38	*dna*E	603	0.41	0.39	0.39
*dna*K	607	0.20	0.68	0.70	*dna*K	611	0.34	0.63	0.66
*dtp*T	463	0.25	0.48	0.45	*yclF*	492	0.46	0.39	0.41
eml	521	0.15	0.61	0.66	*ywk*A	582	0.31	0.60	0.61
*ger*C2	252	0.18	0.45	0.42	*ger*C2	233	0.46	0.38	0.43
*gro*EL	542	0.14	0.59	0.58	*gro*EL	544	0.36	0.67	0.69
*gro*ES	94	0.23	0.48	0.40	*gro*ES	94	0.36	0.57	0.63
*grp*E	179	0.22	0.63	0.49	*grp*E	187	0.39	0.57	0.46
hom	428	0.22	0.41	0.34	*hom*	433	0.48	0.39	0.43
hpt	183	0.14	0.53	0.49	*hprt*	180	0.35	0.42	0.49
*ilv*A	441	0.22	0.41	0.35	*ilv*A	422	0.42	0.39	0.43
*ilv*B	575	0.22	0.43	0.39	*ilv*B	572	0.42	0.38	0.42
*ilv*C	344	0.24	0.53	0.50	*ilv*C	342	0.36	0.53	0.56
*ilv*D	570	0.26	0.45	0.39	*ilv*D	558	0.44	0.43	0.43
*ilv*N	158	0.21	0.36	0.32	*ilv*N	174	0.43	0.37	0.41
*inf*A	72	0.39	0.49	0.40	*inf*A	72	0.35	0.64	0.59
ldh	325	0.23	0.71	0.74	*lct*E	321	0.44	0.42	0.49
*leu*A	513	0.19	0.42	0.38	*leu*A	518	0.39	0.41	0.44
*leu*B	345	0.21	0.38	0.31	*leu*B	365	0.41	0.38	0.39
*leu*C	460	0.21	0.41	0.38	*leu*C	472	0.43	0.42	0.45
*leu*D	191	0.24	0.43	0.40	*leu*D	199	0.38	0.40	0.46
*lmr*A	590	0.25	0.49	0.40	*yvc*C	589	0.47	0.30	0.35
lsp	143	0.23	0.47	0.41	*lsp*	154	0.41	0.38	0.41
*opp*A	600	0.22	0.49	0.43	*opp*A	545	0.38	0.47	0.51
*opp*B	319	0.27	0.34	0.30	*opp*B	311	0.44	0.32	0.35
*opp*C	294	0.26	0.37	0.28	*opp*C	305	0.39	0.36	0.43
*opp*D	338	0.22	0.44	0.34	*opp*D	336	0.46	0.41	0.39
*opp*F	319	0.22	0.39	0.33	*opp*F	307	0.43	0.38	0.39
*pep*P	352	0.25	0.39	0.30	*yqh*T	353	0.46	0.34	0.39
*pep*T	413	0.20	0.58	0.62	*pep*T	410	0.53	0.44	0.41
pfk	340	0.24	0.65	0.66	*pfk*	319	0.37	0.45	0.48
*pgd*H	472	0.24	0.61	0.61	*yqj*L	406	0.38	0.51	0.58
*phe*A	279	0.28	0.35	0.31	*phe*A	285	0.38	0.33	0.38
pyk	502	0.23	0.72	0.78	*pyk*	585	0.37	0.48	0.54
*pyr*D	311	0.28	0.44	0.34	*pyr*D	311	0.43	0.35	0.43
*pyr*F	237	0.22	0.52	0.40	*pyr*F	239	0.42	0.46	0.45
*rec*A	365	0.21	0.60	0.56	*rec*E	347	0.41	0.48	0.49

Table 3-5 (Cont.) Codon usage of *L. lactis* and *B. subtilis* homologues

LL	L	GC_{3s}	F_{op}	CAI	BS	L	GC_{3s}	F_{op}	CAI
*rec*F	357	0.21	0.43	0.37	*rec*F	370	0.47	0.38	0.38
rpmG	48	0.15	0.39	0.33	*rpmG*	49	0.32	0.34	0.51
*rpo*D	340	0.23	0.56	0.54	*rpo*D	371	0.33	0.51	0.50
*rps*M	51	0.14	0.61	0.69	*rps*M	121	0.28	0.72	0.77
*sac*A	473	0.19	0.40	0.34	*sac*A	480	0.48	0.33	0.38
*sec*Y	439	0.26	0.50	0.40	*sec*Y	431	0.39	0.41	0.44
*sod*A	206	0.20	0.57	0.62	*sod*A	226	0.39	0.58	0.63
*thr*B	296	0.27	0.42	0.32	*thr*B	308	0.50	0.35	0.38
tma	695	0.21	0.60	0.57	*fts*H	637	0.45	0.47	0.46
tpi	252	0.24	0.79	0.80	*tpi*	252	0.41	0.59	0.57
*trp*A	253	0.21	0.47	0.39	*trp*A	267	0.45	0.39	0.41
*trp*B	402	0.22	0.47	0.41	*trp*B	400	0.38	0.37	0.43
*trp*C	264	0.27	0.49	0.38	*trp*C	250	0.35	0.42	0.44
*trp*D	335	0.26	0.48	0.37	*trp*D	337	0.41	0.34	0.39
*trp*E	456	0.27	0.44	0.33	*trp*E	515	0.41	0.39	0.41
*trp*F	351	0.22	0.43	0.36	*trp*F	215	0.41	0.35	0.38
*trp*G	198	0.24	0.41	0.35	*trp*G	194	0.46	0.40	0.39
*tyr*A	354	0.21	0.38	0.32	*tyr*A	371	0.39	0.35	0.41
upp	211	0.21	0.64	0.57	*upp*	209	0.37	0.40	0.46

Table 3-5 Codon usage bias in *L. lactis* and *B. subtilis* homologues. Genes which are partial are labelled in blue. Where the function of either gene of a homologous pair has not been identified, the gene has been labelled in red. Table legend: LL is L. lactis gene name; BS is B. subtilis gene name; L is the length of the gene in codons; GC_{3s} is the G+C content at silent positions; F_{op} is the frequency of optimal codons; CAI is the codon adaptation index.

combined codon usage of the two sets of genes was 42.6 for *L. lactis* and 51.0 for *B. subtilis*. This apparently higher bias in *L. lactis* is largely due to the difference in GC_{3s} (*L. lactis* GC_{3s}=0.23, *B. subtilis* GC_{3s}=0.41) between the two sets of genes. Overall, *L. lactis* genes have a higher F_{op} than their *B. subtilis* homologues (see Figure 3-5), this is despite *B. subtilis* having a greater number of defined optimal codons (19 optimal codons, compared with 17 for *L. lactis*). There is a significant correlation between the F_{op} values of both sets of homologous genes (r=0.637, p<0.001), the relationship between these two sets of F_{op} values is presented in Figure 3-5.

Relative to their F_{op} values in *L. lactis, gro*ES, *gro*EL, *aro*A, *inf*A, and *rps*M F_{op} are higher in *B. subtilis*. For *rpm*G this result is deceptive, as the *L. lactis* sequence analyzed is a partial 5' sequence, and when F_{op} is calculated for the first 51 codons of the *B. subtilis rpm*G sequence both values are identical (F_{op}=0.61). This reflects a general trend in *L. lactis* chromosomal sequences where the F_{op} of the first 50 codons is lower than the remainder of the gene and is most pronounced in highly expressed genes (data not shown). This variation in codon usage along a gene has been reported for *E. coli* (Eyre-Walker and Bulmer 1993), but was not seen in *B. subtilis* (Sharp 1990) emphasises a potential difficulty when analysing 5' partial sequences.

Interestingly while the heat shock proteins HSPs *gro*ES and *gro*EL have a higher F_{op} in *B. subtilis*, the HSPs *grp*E, and *dna*K have a higher F_{op} in *L. lactis*. Selection for optimal codons in *gro*ES, *gro*EL, *aro*A, and *inf*A appears to be more effective in *B. subtilis* than in *L. lactis*.

While the average F_{op} is higher in *L. lactis,* the genes *lmr*A*, adk, upp, pyk, pfk, ldh* and *tpi*, have markedly increased F_{op} value. The *tpi*, *pyk, pfk, ldh*, and *adk* genes encode key glycolytic enzymes. When one considers that many of the common *L. lactis* strains are used in rapid homolactic fermentations, it is not unexpected that selection for optimal codons is successful in these genes (van de Guchte, Kok and Venema 1992). The *B. subtilis* homologue (*yvc*C) of the *L. lactis* gene *lmr*A was identified on the basis of sequence identity, therefore the most simple explanation for this increase in F_{op} is that this gene pair are not homologues. Another gene with a markedly increased F_{op} was *upp* which encodes the major pathway for uracil recycling

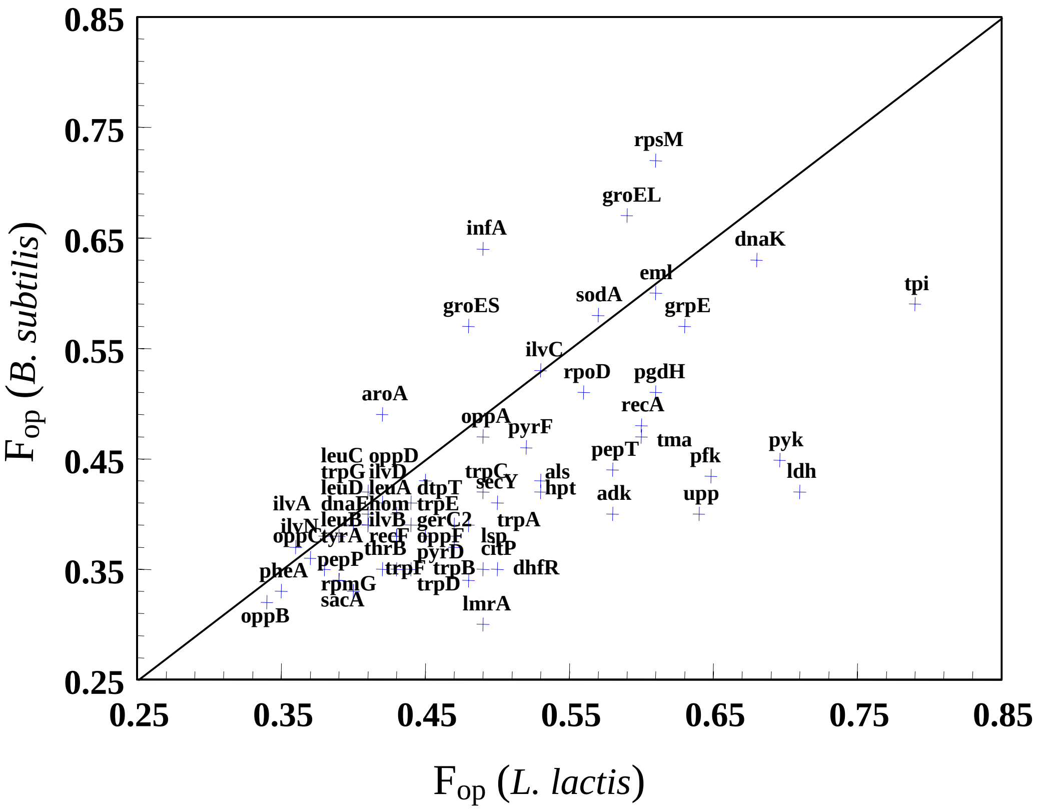

Figure 3-5. Correlation between the frequency of optimal codons in 73 homologous *B. subtilis* and *L. lactis* genes.

(Martinussen and Hammer 1994), an key pathway in RNA metabolism and which may have a role in overall fitness.

3.4.2.6 L. lactis p-aminobenzoic acid synthetase

The *L. lactis* p-aminobenzoic acid synthetase gene *pabB* (Arhin and Vining 1993); (GenBank accession no. M64860) was removed from the original dataset because it had high codon bias but atypical codon usage. COA of 124 *L. lactis* genes including *pab*B clearly identified it as an outlier (see Figure 3-6). Having identified optimal *L. lactis* codons, these can be used to contrast the codon usage of *pabB* with other *L. lactis* genes. While it is good practice to remove anomalous genes from a COA dataset, the inclusion of *pab*B had little affect on the principal axis, and did not change the optimal and non optimal codons, but it had a marked influence on the 2nd and subsequent correspondence analysis axes (see Figure 3-6). Alignments of the amino acid sequence of the *L. lactis pabB* gene with other *pabB* genes confirms that it has significant sequence similarity, 38% identity with *E. coli* and 48% identity with *Streptomyces lividans*.

The gene *pabB* is 471 codons in length and has an extremely biased codon usage with an EN_c of 22.2; the next lowest EN_c in this dataset was 27.2 for *nrd*H. Despite having a strong bias in codon usage their F_{op} (*pab*B=0.45, *nrd*H=0.51) were atypical, being mid-ranged compared with other *L. lactis* sequences, when EN_c is plotted against F_{op} these two sequences are outliers (see Figure 3-7). It has already been noted that *nrd*H is probably too short to accurately calculated ENc but pabB exceeds the recommend 100 codon limit four fold. Of the 17 optimal codons identified in *L. lactis*, 6 are unused (CGU, AAA, CAC, UAC and CUU) and two universal optimal codons (GGU and GAC) are used only twice, despite the amino acids Gly (N=54) and Asp (N=29) being relatively common. The difference between the codon usage of *pabB* and other *L. lactis* sequences can be seen in its very bias codon usage (see Table 3-6) and by contrasting the codon usage of the amino acids Lys and Arg (see Table 3-7). The bias is very extreme but does not appear to follow any logical patter, it does not seem to be a mutational bias as there is almost complete usage of

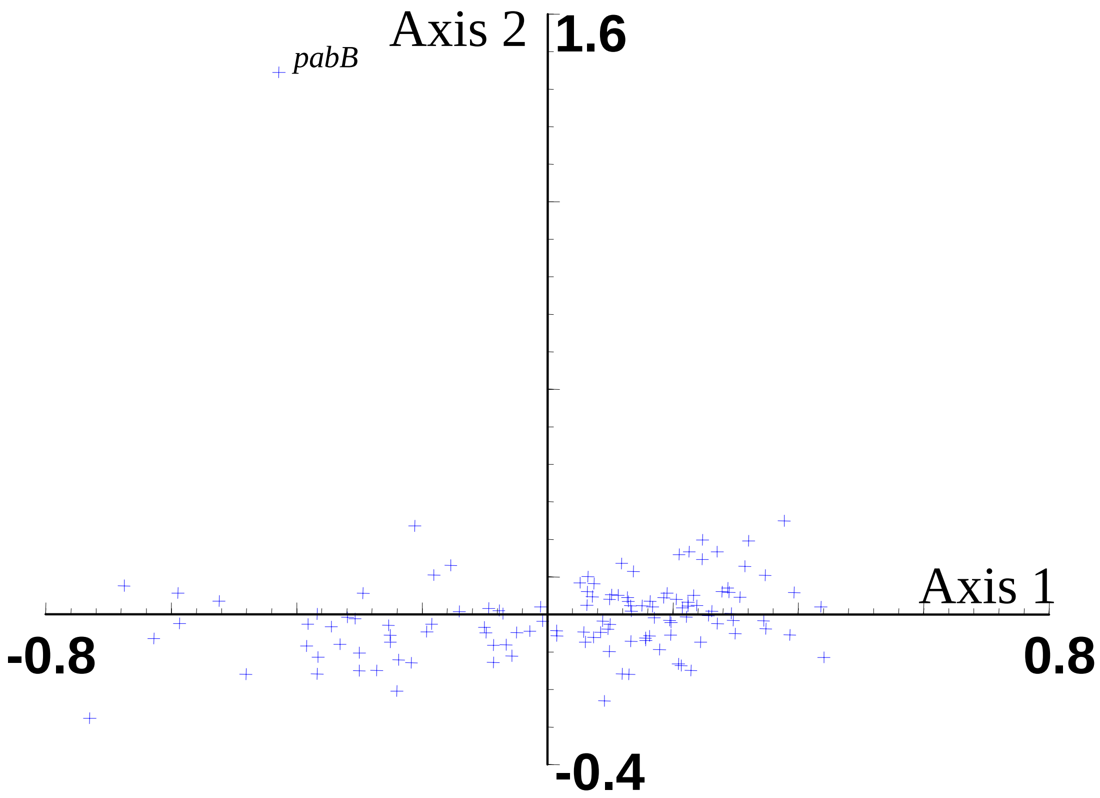

Figure 3-6 Correspondence analysis of 124 L. lactis genes (including pabB), the abnormal codon usage of pabB is demonstrated by its position on Axis2.

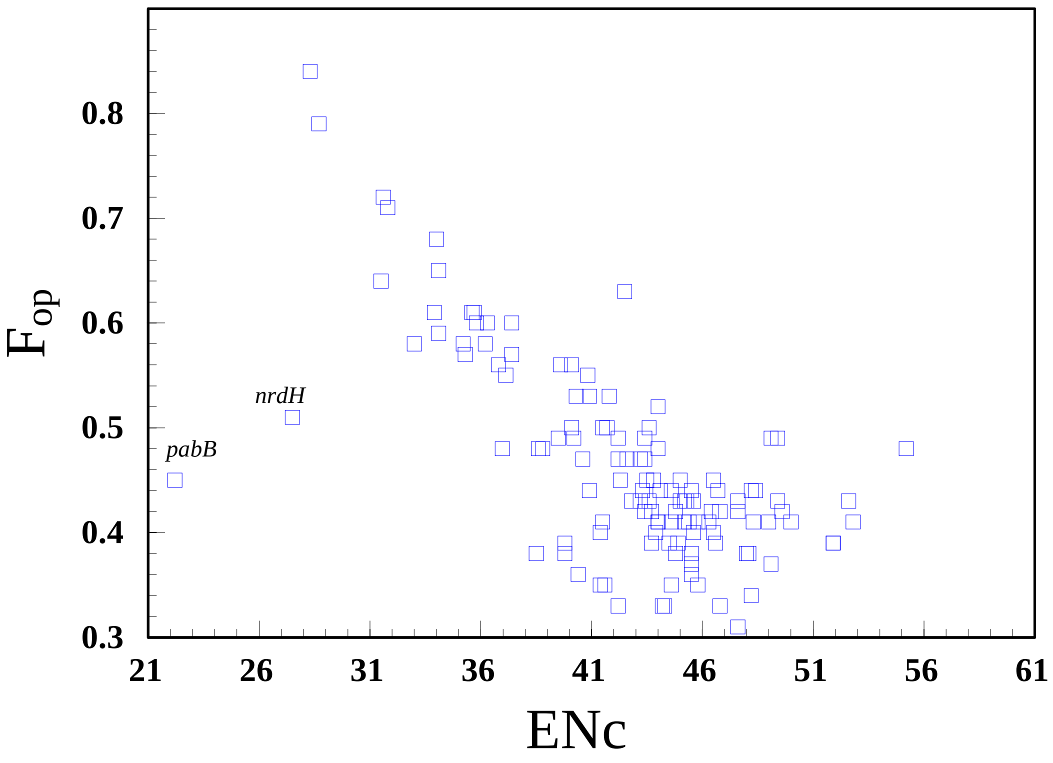

Figure 3-7. Correlation between EN_c and F_{op} for *L. lactis* genes. *pabB* and *hrdH* have unusually low F_{op} values compared with their EN_c values.

		N	RSCU			N	RSCU			N	RSCU			N	RSCU
Phe	UUU	0	0.00	Ser	UCU	0	0.00	Tyr	UAU	0	0.00	Cys	UGU	8	2.00
	UUC	17	2.00		UCC	1	0.22		UAC	20	2.00		UGC	0	0.00
Leu	UUA	0	0.00		UCA	0	0.00	TER	UAA	1	3.00	TER	UGA	0	0.00
	UUG	34	5.83		UCG	0	0.00		UAG	0	0.00	Trp	UGG	2	1.00
	CUU	0	0.00	Pro	CCU	3	0.80	His	CAU	8	2.00	Arg	CGU	0	0.00
	CUC	0	0.00		CCC	0	0.00		CAC	0	0.00		CGC	0	0.00
	CUA	0	0.00		CCA	11	2.93	Gln	CAA	14	1.87		CGA	2	0.46
	CUG	1	0.17		CCG	1	0.27		CAG	1	0.13		CGG	0	0.00
Ile	AUU	0	0.00	Thr	ACU	6	1.33	Asn	AAU	1	0.07	Ser	AGU	26	5.78
	AUC	32	2.91		ACC	0	0.00		AAC	29	1.93		AGC	0	0.00
	AUA	1	0.09		ACA	12	2.67	Lys	AAA	0	0.00	Arg	AGA	24	5.54
Met	AUG	18	1.00		ACG	0	0.00		AAG	8	2.00		AGG	0	0.00
Val	GUU	0	0.00	Ala	GCU	2	0.19	Asp	GAU	28	1.93	Gly	GGU	1	0.07
	GUC	0	0.00		GCC	0	0.00		GAC	1	0.07		GGC	0	0.00
	GUA	33	3.88		GCA	39	3.71	Glu	GAA	30	1.94		GGA	53	3.93
	GUG	1	0.12		GCG	1	0.10		GAG	1	0.06		GGG	0	0.00

Table 3-6 Codon usage of *pabB* a gene with an extreme codon bias

		High bias		Low bias		*pabB*	
Lys	AAA	1.90	(93)	1.63	(120)	0	(0)
	AAG	0.10	(5)	0.37	(27)	1.0	(8)
Arg	CGU	5.32	(55)	1.00	(9)	0	(0)
	CGC	0.68	(7)	0.44	(4)	0	(0)
	CGA	0.00	(0)	1.11	(10)	0.48	(2)
	CGG	0.00	(0)	0.44	(4)	0	(0)
	AGA	0.00	(0)	2.56	(23)	5.52	(24)
	AGG	0.00	(0)	0.44	(4)	0	(0)

Table 3-7 Comparison of the codon usage of the *pabB* gene with groups of other *L. lactis* genes. The highly and lowly biased codon usage data is taken from Table 3-4. RSCU values are presented, the values in parenthesis are codon frequencies.

A-ending codons for some amino acids (Val, Ala, Glu, Gly, Arg, Gln, Pro, and Thr), C-ending for others (Phe, Tyr, Ile and Asn) and G-ending for others (Phe, Cys, His, and Ser), and it would require a very extreme selection pressure to account for such a high observed bias. The codons AAG and AGA are both "rare codons" and are used infrequently in *L. lactis*, yet *pabB* preferentially uses AGA for Arginine (RSCU=5.52, N=24) and AAA for Lysine (RSCU=2.0, N=8). Chi-squared tests on the usage of Lys and Arg codons confirm that these differences are highly significant (all tests p<0.001).

If this sequence is not fictitious, which on the face of it seems to be the most plausible explanation, then it must have been recently acquired by horizontal gene transfer from a species which has a codon usage unlike anything so far reported.

The discovery of *pab*B's unusual codon usage demonstrates the difference between a detailed analysis of codon usage and the simple tabulation of codon usage. Unless the codon usage, especially the frequency of optimal and non-optimal codons are examined it is not obvious that this gene is anomalous. This sequence was included by Llanos *et al*. 1993 as an example of a ordinary chromosomal gene in their analysis of codon usage bias in the *las* operon and as dataset only contained 26 other genes it had a noticeable influence on the frequency of codons in their tabulation. While only 4% of chromosomal codons where from *pab*B, it accounted for 22% of AGA, 16% of GGA, and 15% of AGU, codons in their table labelled "Codon usage of *L. lactis* genes". The original authors of the sequence did not identify anything unusual about its codon usage (Arhin and Vining 1993).

3.4.2.7 Translation Termination Codon Usage

Among the 117 termination codons in the dataset the most frequent stop codon was TAA (70%) followed by TGA (22%) and TAG (8%). This usage pattern is similar to that found in *E. coli* and *B. subtilis,* where the most common stop codon is TAA and the least common is TAG (Lloyd and Sharp 1992b). The nucleotide adenine is the most common nucleotide (39%) after a stop codon, followed by thymidine (35%), guanine (16%), and cytosine (10%). The tri-nucleotides, AAA and TTT, are found significantly more frequently than expected (AAA N=13 p<0.01; TTT N=11 p<0.01) immediately 3' of the termination codon. In many species, including *S. cerevisiae, E. coli*, and *B. subtilis*, the choice of termination codon is correlated with gene expression level (Sharp and Bulmer 1988). A similar trend is seen in *L. lactis* when the usage of termination codons is compared between genes of different predicted expression levels. In Figure 3-8, where the genes have been split into 5 groups on the basis of their F_{op}, the stop codon TAA is used exclusively in the 23 genes with the highest F_{op}. In addition to the preference for the stop codon TAA amongst these 20 highly expressed genes, there is a strong preference for the following nucleotide to be thymidine (N=14), and for the following trinucleotide to be either TTT (N=8) or TAA (N=3). These results suggest that the translation termination signal extends beyond the stop codon triplet and that the efficiency of translation termination signal is correlated with gene expression level.

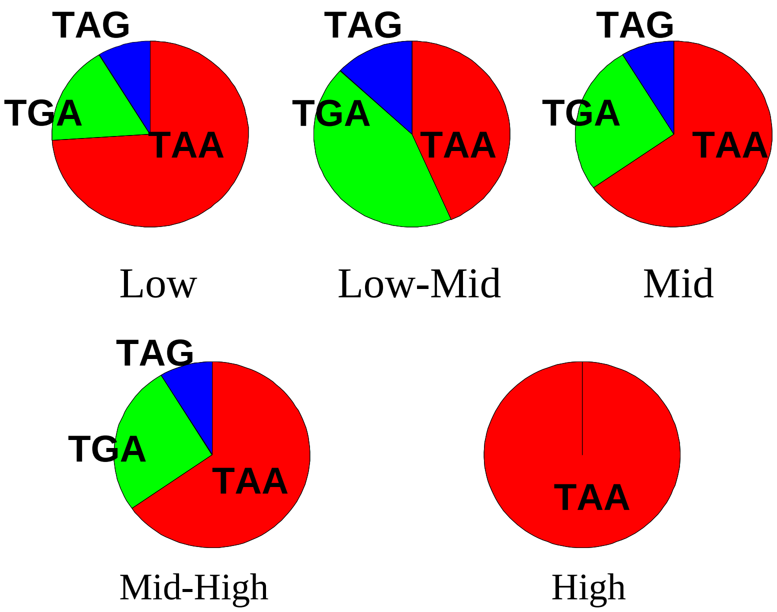

Figure 3-8. Translation termination codon usage. The F_{op} of 117 *L. lactis* genes, where the choice of termination codon was also known, was used to rank and separate the gene sequences into five groups, approximately equal in size. The groups were labelled according to their F_{op}, High (highest Fop), Med-High, Medium, Low-Med, Low. The usage of each of the three termination codons within each group is displayed as a pie chart.

4 *Staphylococcus aureus* and *Streptococcus mutans*

4.1 Staphylococcus aureus

Natural populations of *Staphylococcus* are mainly associated with the skin, skin glands and mucous membranes of warm-blooded animals. *Staphylococcus aureus*, the type organism for the genus, is a Gram-positive coccus occurring singly or in pairs, in which division occurs in more than one plane, giving rise to characteristic clusters. It is a facultative anaerobe with an overall G+C content of 32-36%, phylogenetically related to *B. subtilis* (see Figure 4-1). It is a pathogen in a wide range of infections including furuncles, carbuncles, wound infections, toxic shock syndrome, food poisoning (via enterotoxins), and mastitis in man and domestic animals. Most strains possess the species-specific protein A, surface-bound and secretory coagulase and DNAse. Acid is produced aerobically and anaerobically by most strains when grown with lactose as a sole carbohydrate source. At least four different exotoxins (α- β- γ- and δ-hemolysins) are produced, with some strains also producing bacteriocins. DNA/DNA hybridisation studies of strains of *S. aureus* have shown that they have not diverged by more than 3%, but they have also confirmed that *S. aureus* is not closely related to other *Staphylococcus* species (Kloos and Schleifer 1986).

4.1.1 Analysis of Codon usage of *S. aureus*.

The codon usage of *S. aureus* was examined using a similar protocol to that described for *L. lactis*. All annotated coding sequences for both species were extracted from GenBank release 95. This produced a 739 sequence dataset for *S. aureus*, those sequences which were partial or likely to have been horizontally transferred using the criteria described above were removed. The *lac* operon genes were also removed. There was an unusually large amount of sequence redundancy in the original dataset due to the presence of numerous copies of sequences associated with strain-specific virulence determinants of *S. aureus*. This process reduced the dataset to 179 genes. The total codon usage of these genes is tabulated in Table 4-1a. As expected for a Low G+C species there is a predominance of A/U ending codons.

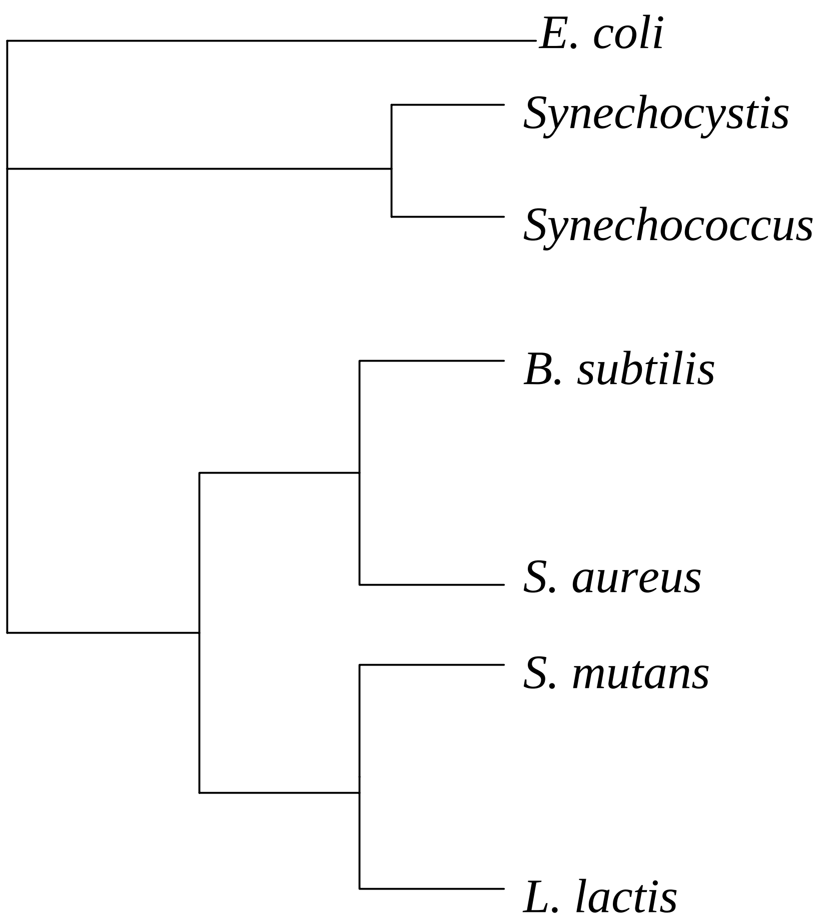

Figure 4-1 Diagrammatic representation of the relationship between species analysed in this thesis and *E. coli* and *B. subtilis***.**

Table 4-1 Overall codon usage

a) Total codon usage of 179 *S. aureus* genes (67876 codons)

		N	RSCU			N	RSCU			N	RSCU			N	RSCU
Phe	UUU	1994	1.42	Ser	UCU	982	1.39	Tyr	UAU	2154	1.55	Cys	UGU	242	1.53
	UUC	819	0.58		UCC	93	0.13		UAC	634	0.45		UGC	74	0.47
Leu	UUA	3323	3.55		UCA	1373	1.95	TER	UAA	128	2.15	TER	UGA	20	0.34
	UUG	807	0.86		UCG	242	0.34		UAG	31	0.52	Trp	UGG	500	1.00
	CUU	713	0.76	Pro	CCU	837	1.40	His	CAU	1041	1.53	Arg	CGU	979	2.38
	CUC	99	0.11		CCC	57	0.10		CAC	324	0.47		CGC	165	0.40
	CUA	529	0.56		CCA	1203	2.01	Gln	CAA	2215	1.77		CGA	276	0.67
	CUG	147	0.16		CCG	293	0.49		CAG	286	0.23		CGG	42	0.10
Ile	AUU	3230	1.81	Thr	ACU	1274	1.24	Asn	AAU	3137	1.46	Ser	AGU	1130	1.60
	AUC	894	0.50		ACC	153	0.15		AAC	1161	0.54		AGC	405	0.58
	AUA	1230	0.69		ACA	2112	2.06	Lys	AAA	4666	1.65	Arg	AGA	872	2.12
Met	AUG	1526	1.00		ACG	561	0.55		AAG	1001	0.35		AGG	137	0.33
Val	GUU	1888	1.62	Ala	GCU	1340	1.36	Asp	GAU	3072	1.52	Gly	GGU	2349	2.16
	GUC	421	0.36		GCC	249	0.25		GAC	968	0.48		GGC	627	0.58
	GUA	1685	1.45		GCA	1844	1.87	Glu	GAA	4019	1.69		GGA	1098	1.01
	GUG	670	0.57		GCG	509	0.52		GAG	748	0.31		GGG	278	0.26

b) Total codon usage of 57 *S. mutans* genes (25100 codons)

		N	RSCU			N	RSCU			N	RSCU			N	RSCU
Phe	UUU	716	1.45	Ser	UCU	402	1.55	Tyr	UAU	802	1.49	Cys	UGU	61	1.39
	UUC	273	0.55		UCC	122	0.47		UAC	276	0.51		UGC	27	0.61
Leu	UUA	526	1.53		UCA	398	1.53	TER	UAA	35	1.84	TER	UGA	11	0.58
	UUG	471	1.37		UCG	78	0.30		UAG	11	0.58	Trp	UGG	275	1.00
	CUU	580	1.69	Pro	CCU	348	1.61	His	CAU	346	1.45	Arg	CGU	464	3.07
	CUC	173	0.50		CCC	74	0.34		CAC	132	0.55		CGC	168	1.11
	CUA	135	0.39		CCA	331	1.53	Gln	CAA	734	1.38		CGA	58	0.38
	CUG	177	0.52		CCG	113	0.52		CAG	327	0.62		CGG	48	0.32
Ile	AUU	1134	2.09	Thr	ACU	551	1.36	Asn	AAU	1106	1.50	Ser	AGU	364	1.40
	AUC	381	0.70		ACC	247	0.61		AAC	373	0.50		AGC	192	0.74
	AUA	110	0.20		ACA	601	1.48	Lys	AAA	1202	1.37	Arg	AGA	130	0.86
Met	AUG	541	1.00		ACG	220	0.54		AAG	554	0.63		AGG	39	0.26
Val	GUU	886	2.05	Ala	GCU	1054	2.00	Asp	GAU	1267	1.56	Gly	GGU	787	1.88
	GUC	344	0.79		GCC	331	0.63		GAC	359	0.44		GGC	369	0.88
	GUA	277	0.64		GCA	551	1.05	Glu	GAA	1240	1.63		GGA	377	0.90
	GUG	226	0.52		GCG	171	0.32		GAG	279	0.37		GGG	145	0.35

N is codon frequency, RSCU is relative synonymous codon usage.

A correspondence analysis of the codon usage of these genes found that the principal factor in the variation in the codon usage was sufficient to explain 17% of the total variation. The second most important factor explained 8.9% of the remaining variation. The projection of the genes on the principal axis (see Figure 4-2 and Table 4-2) was examined for evidence to support the hypothesis that the most important trend in the codon usage of *S. aureus* gene sequences is selection for translational efficiency. It was observed that when the genes are sorted according to their principal axis coordinates, (see Table 4-2) ribosomal protein genes, heat shock proteins and metabolic enzymes predominate at one end, while signal, capsular biosynthesis, membrane transport and regulatory proteins predominate at the other. It was also observed that the EN_c index was significantly correlated (r= 0.55, p<0.001) with the principal axis (axis 1). Those genes with negative coordinates on the principal axis have a more biased usage of codons compared to genes with positive axis 1 coordinates. Both these observations are typical for species in which selection for translational efficiency has been shown to bias codon usage towards a set of optimal codons.

The codon usage of two groups of nine genes (5% of the dataset) from both ends of the principal axis were compared to identify codons which could be putatively described as optimal, i.e. those codons used significantly more often in highly expressed genes and to identify putative rare codons (see Table 4-3). A total of 15 putative optimal codons were identified for 15 amino acids. No optimal codons were identified for Tyr, His, or Cys. These optimal codons included five of the six universally optimal codons (i.e. UUC, AUC, GAC AAC, and GGU). The sixth universal codon UAC (Tyr), while used more frequently in the high bias dataset was only significant at p<0.05. Eleven codons were classified as rare including CGG, AGG, GGG, and CCC which are also considered to be rare in *E. coli* and *B. subtilis* (see Table 4-6). Overall the putative optimal and rare codons of *S. aureus* share a strikingly similarity to the codon usage of *B. subtilis*, *E. coli*, and *L. lactis*. This overlap between the putative optimal codons of *S. aureus* and those identified as optimal in other species, in conjunction with the functional role of genes using these preferred codons, is compelling evidence that there is selection for a subset of preferred codons in *S. aureus* and that these codons are translationally optimal.

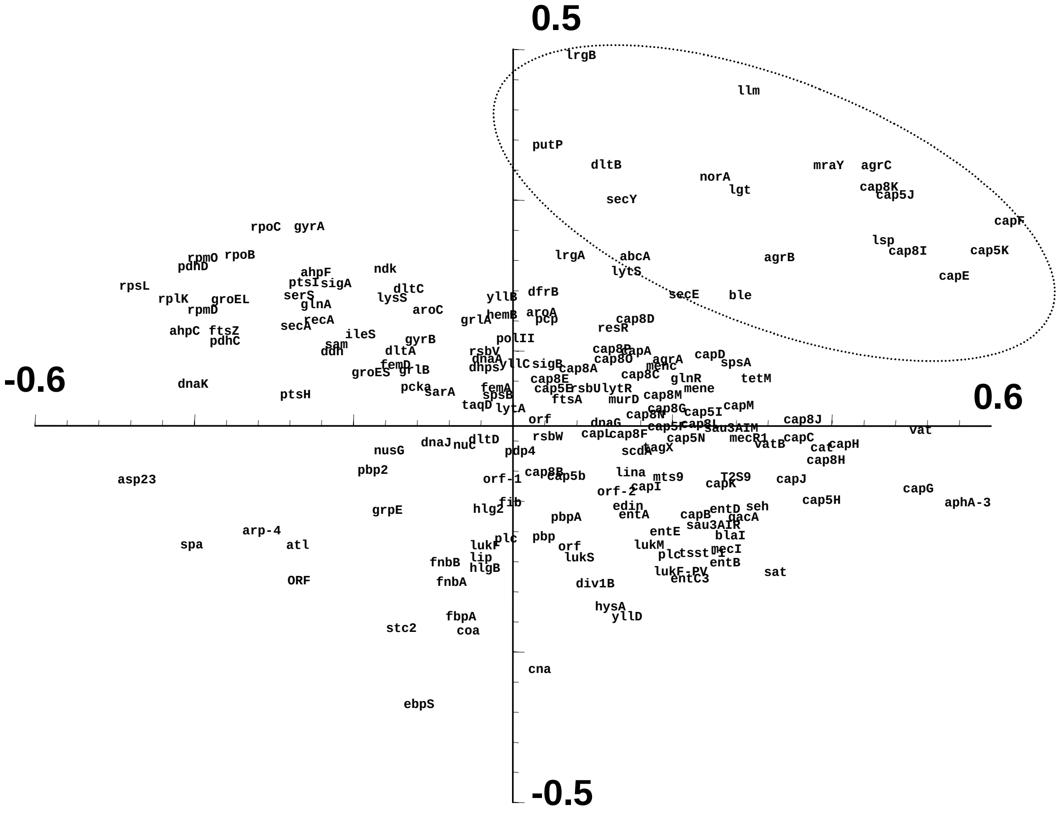

Figure 4-2 Correspondence analysis of the codon usage of 179 *S. aureus* genes. Each gene is plotted using its gene name at its coordinate on the first two axes produced by the analysis. The putatively highly expressed genes lie towards the left side of axis 1. The dashed ellipse delineates a group of putative membrane proteins.

Table 4-2 ***S. aureus*** gene sequences

Gene	GC_{3s}	L	Gene description	ENc	Fop	Acc.#	Pid	Reference:
asp23	0.13	169	alkaline shock protein 23	30.59	0.73	S76213	g894289	ZMEI 3:28
rpsL	0.20	137	ribosomal protein S12	34.11	0.64	U20869	g706921	Unpublished
rplK	0.16	140	ribosomal protein L11	34.79	0.63	U96619	g2078378	Unpublished
ahpC	0.21	189	alkyl hydroperoxide reductase subunit C	32.93	0.64	U92441	g1916316	Micro 141:1655
pdhD	0.15	468	dihydrolipoamide dehydro: subunit E3	33.69	0.63	X58434	g48874	BBA 1129: 119
dnaK	0.16	610	heat shock protein 70	34.90	0.61	D30690	g441211	JBa 176:4779
spa	0.27	454	protein A	36.70	0.61	X61307	g46691	FEMS 91:1
rpmD	0.17	59	ribosomal protein L30	27.88	0.62	U96620	g2078380	Unpublished
rpmO	0.15	146	ribosomal protein L15	39.51	0.63	U96620	g2078381	Unpublished
ftsZ	0.16	390	cell wall division	35.31	0.55	U06462	g458428	Micro 144:3069
pdhC	0.15	430	dihydrolipoamide acetyltransferase: subunit E2	34.36	0.53	X58434	g581570	BBA 1129:119
groEL	0.14	539	heat shock protein 60	34.24	0.53	D14711	g441208	CMBR 193:730
rpoB	0.19	1182	DNA-directed RNA polymerase beta chain	35.99	0.53	X64172	g677851	BBA 1262:73
arp-4	0.24	386	immunoglobulin A binding	44.50	0.52	A09523	g412259	Pat:0367890-A
rpoC	0.17	541	DNA directed RNA polymerase beta' chain	36.55	0.55	Y09428	g1684751	Unpublished
ptsH	0.21	88	histidin-containing protein	38.59	0.56	X93205	g1070385	Unpublished
secA	0.19	843	secretion	36.10	0.50	U97062	g2078390	Unpublished
serS	0.12	428	seryl-trna synthetase	31.63	0.50	Y09924	e291101	BBA1397:169
atl	0.17	1256	autolysin	34.84	0.52	L41499	g765073	JBa 177:5723
ORF	0.22	357	glutamic acid specific protease	37.89	0.46	D00730	g216971	BBA 1121:221
ptsI	0.17	572	phosphoenolpyruvate-protein phosphatase	37.43	0.52	X93205	g1070386	Unpublished
glnA	0.17	446	glutamine synthetase	35.03	0.52	X76490	g1134886	JBa 176:1460
gyrA	0.17	886	DNA gyrase	36.26	0.48	X71437	g296396	JBa 175:3269
ahpF	0.23	507	alkyl hydroperoxide reductase subunit F	37.60	0.48	U92441	g1916317	Micro 141:1655
recA	0.15	347	genetic recombination	39.09	0.51	L25893	g463285	Gene 147:13
ddh	0.18	330	D-specific D-2-hydroxyacid dehydrogenase	35.27	0.48	U31175	g1644433	AAC 40:166
sigA	0.13	368	sigma 70	34.87	0.50	AB001896	g1943995	Unpublished
sam	0.18	397	S-adenosylmethionineynthetase	38.17	0.42	U36379	g1020317	Unpublished
ileS	0.17	917	isoleucyl-tRNA synthetase	35.20	0.49	X74219	g437916	Gene 141:103
groES	0.14	94	heat shock protein 10	41.89	0.50	D14711	g441207	CMBR 193:730
pbp2	0.20	716	penicillin-binding protein 2	38.17	0.47	X62288	g483534	FEMS 117:131
grpE	0.14	208	heat shock protein 20	39.86	0.51	D30690	g441210	JBa 176:4779
nusG	0.18	182	transcription antitermination	37.58	0.49	U96619	g2078377	Unpublished
ndk	0.10	149	nucleoside diphosphate kinase	32.51	0.49	U31979	g987497	Micro 142:2943
lysS	0.17	495	lysyl-tRNA synthetase	38.16	0.47	L36472	g567884	Unpublished
femD	0.13	451	phosphoglucomutase	34.73	0.48	Y09570	g1684749	Unpublished
dltA	0.18	337	D-alanine-D-alanyl protein ligase	38.04	0.46	D86240	g1405335	Unpublished
stc2	0.23	715	staphylocoagulase precursor	38.94	0.42	D00184	g216977	JBio 102:1177
dltC	0.11	78	D-alanyl carrier protein	25.27	0.45	D86240	g1405337	Unpublished
grlB	0.17	663	DNA topoisomerase IV GrlB subunit	38.78	0.46	D67074	g1777317	AAC 40:1157
pckA	0.22	530	phosphoenolpyruvate carboxykinase	40.25	0.46	U51133	g1255262	Unpublished
ebpS	0.19	202	elastin binding protein	36.70	0.41	U48826	g1397239	JBC 271:15803
gyrB	0.22	640	DNA gyrase B subunit	41.26	0.40	M86227	g153085	JBa 174:1596

Table 4-2 ***S. aureus*** gene sequences (cont.)

Gene	GC_{3s}	L	Gene description	ENc	Fop	Acc.#	Pid	Reference:
aroC	0.16	388	chorismate synthase	37.22	0.37	U31979	g987498	Micro 142:2943
dnaJ	0.16	379	heat shock protein 40	36.76	0.43	D30690	g522106	JBa 176:4779
sarA	0.21	124	regulator A of agr expression	42.55	0.50	U20782	g684950	JBa 176:4168
fnbB	0.20	940	fibronectin binding protein B	40.19	0.44	X62992	g581562	EJB 202:1041
fnbA	0.19	1018	fibronectin-binding protein	39.60	0.42	J04151	g295152	PNAS 86:699
fbpA	0.18	645	fibrinogen binding protein	38.14	0.41	U20794	g915308	IIm 63:1914
nuc	0.19	231	nuclease	38.84	0.46	V01281	g673492	Gene 22:181
coa	0.20	636	coagulase	40.54	0.39	X17679	g46540	MM 4:393
grlA	0.16	800	DNA topoisomerase IV GrlA subunit	36.51	0.44	D67074	g1777318	AAC 40:1157
taqD	0.19	132	glycerol-3-phosphate cytidyltransferase	45.23	0.42	X87105	g1125683	AAC 39:2415
dltD	0.18	254	extramembranal protein	38.15	0.44	D86240	g1405338	Unpublished
rsbV	0.15	108	alternative sigma factor	35.41	0.44	Y07645	g1934989	AMic 167:151
dhps	0.17	267	dihydropteroate synthase	40.66	0.40	Z84573	g2058356	JMB (In press)
lip	0.19	682	glycerol ester hydrolase	40.44	0.40	M90693	g393266	Unpublished
lukF	0.18	323	leukocidin F component	38.21	0.42	S65052	g410007	Unpublished
dnaA	0.17	453	chromosomal replication initiator	39.95	0.43	D89066	g1694677	MGG 246:680
hlgB	0.18	325	gamma-hemolysin component B	37.66	0.41	L01055	g295156	JGM 134:2179
hlg2	0.21	309	gamma-hemolysin II	38.53	0.43	S65052	g410005	Unpublished
femA	0.19	433	cell wall or membrane metabolism	39.07	0.44	X17688	g46581	MGG 219:263
spsB	0.16	191	type-I signal peptidase SpsB	37.80	0.44	U65000	g1595810	JBa 178: 5712
orf-1	0.19	240	novel antigen	36.20	0.37	U60589	g1407784	Unpublished
yllB	0.25	144	cell wall division	44.59	0.41	U94706	g2149890	Unpublished
hemB	0.17	323	porphobilinogen synthase	39.02	0.39	S72488	g632816	BBB 57:1234
yllC	0.23	311	cell wall division	42.05	0.41	U94706	g2149891	Unpublished
plc	0.21	331	beta-hemolysin	47.08	0.42	S72497	g619317	CJM 40:651
lytA	0.30	481	peptidoglycan hydrolase	48.57	0.38	M76714	g153067	Gene 102:105
polII	0.20	1415	DNA polymerase III	40.03	0.39	Z48003	g642270	Gene 165:51
fib	0.20	165	fibrinogen-binding protein	37.39	0.40	X72013	g311974	MM 12599
pdp4	0.21	431	penicillin binding protein 4	42.66	0.38	X87105	g1125682	AAC 39:2415
cap8B	0.21	228	capsular polysacc. biosynthesis	42.00	0.34	U73374	g1657641	JBa 179:1614
aroA	0.18	430	3-phosphoshikimate-1-carboxyvinyltransferase	40.77	0.35	L05004	g152956	JGM 139:1449
dfrB	0.17	159	dihydrofolate reductase	40.91	0.39	Z16422	g49313	AAC 37:1400
cna	0.24	1183	collagen adhesin	39.04	0.38	M81736	g387880	JBC 267:4766
URF	0.22	419	unknown	38.89	0.40	X17688	g46582	MGG 219:263
cap8E	0.20	342	capsular polysacc. biosynthesis	40.74	0.35	U73374	g1657644	JBa 179:1614
sigB	0.19	256	sigma factor B	44.10	0.41	Y07645	g1934991	AMic 167:151
putP	0.20	497	proline permease homolog	39.33	0.38	U06451	g458420	AEM 61:252
rsbW	0.21	159	alternative sigma factor	39.44	0.36	Y07645	g1934990	AMic 167:151
pbp	0.20	670	penicillin-binding protein	42.32	0.42	Y00688	g46629	Gene 94:137
cap5E	0.20	342	capsular polysacc. biosynthesis	39.99	0.35	U81973	g1773344	Unpublished
pcp	0.19	212	pyrrolidone carboxyl peptidase	43.76	0.30	U19770	g790573	Gene 166:95
ftsA	0.20	471	cell division protein	44.17	0.38	U94706	g2149897	Unpublished
cap5b	0.20	228	Cap5B capsid protein	42.36	0.32	U81973	g1773341	Unpublished
pbpA	0.19	744	penicillin-binding protein 1	41.75	0.38	U94706	g2149893	Unpublished
rsbU	0.20	333	cell wall division	41.21	0.41	Y09929	g1729794	JBa 178: 6036
lrgA	0.24	147	holin-like protein LrgA	42.30	0.36	U52961	g1575025	JBa 178:5810
cap8A	0.23	222	capsid protein	41.56	0.42	U73374	g1657640	JBa 179:1614
ORF	0.20	689	human MHC class II analog	37.87	0.40	U20503	g1001961	JBC 270:21457

Table 4-2 ***S. aureus*** gene sequences (cont.)

Gene	GC$_{3s}$	L	Gene description	ENc	Fop	Acc.#	Pid	Reference:
lukS	0.23	315	leukocidin S component	39.54	0.36	S65052	g410006	Unpublished
lrgB	0.25	233	similar to *E.coli* yohK	40.80	0.38	U52961	g1575026	JBa 178:5810
div1B	0.21	439	cell division protein	42.05	0.38	U94706	g2149896	Unpublished
capL	0.17	424	capsular polysacc. biosynthesis	37.20	0.35	U10927	g506708	JBa 176:7005
dnaG	0.19	572	DNA primase	38.74	0.35	AB001896	g1943994	Unpublished
dltB	0.23	404	predicted membrane transporter	35.03	0.41	D86240	g1405336	Unpublished
cap8P	0.22	391	capsular polysacc. biosynthesis	42.99	0.27	U73374	g1657655	JBa 179:1614
cap8O	0.21	420	capsular polysacc. biosynthesis	41.16	0.34	U73374	g1657654	JBa 179:1614
hysA	0.25	807	hyaluronate lyase	45.43	0.38	U21221	g705406	FEMS 130:81
orf-2	0.20	239	novel antigen	44.69	0.33	U63529	g1488695	Unpublished
resR	0.14	192	DNA invertase	38.32	0.35	M36694	g153027	FEMS 50:253
lytR	0.22	246	affects autolysis	41.84	0.32	L42945	g1854577	JBac78:611
secY	0.22	430	membrane transport	39.72	0.40	U96620	g2078382	Unpublished
murD	0.19	450	D-glutamic acid adding enzyme	42.22	0.34	U94706	g2149895	Unpublished
cap8F	0.24	371	capsular polysacc. biosynthesis	45.41	0.30	U73374	g1657645	JBa 179:1614
yllD	0.24	133	homologous to ftsL	42.10	0.38	U94706	g2149892	Unpublished
lytS	0.25	584	affects autolysis	44.96	0.32	L42945	g862312	JBa178:611
edin	0.14	247	epidermal cell differentiation inhibitor	37.02	0.39	M63917	g152998	BBRC 174:459
lina	0.11	161	lincosaminide nucleotidyltransferase	35.88	0.44	J03947	g153041	JBC 263:15880
cap8D	0.24	607	capsid protein	46.12	0.33	U73374	g1657643	JBa 179:1614
entA	0.14	280	enterotoxin D precursor	32.79	0.38	M17347	g153006	JBa 169:3904
abcA	0.20	575	ATP-binding cassette transporter	41.66	0.37	X91786	g1262136	AAC 40:2121
cap5F	0.26	371	capsular polysacc. biosynthesis	47.09	0.30	U81973	g1773345	Unpublished
capA	0.11	221	capsular polysacc. biosynthesis	36.77	0.44	U10927	g506697	JBa 176:7005
cap8C	0.21	254	capsular polysacc. biosynthesis	40.80	0.36	U73374	g1657642	JBa 179:1614
scdA	0.20	224	cell wall division	40.14	0.32	U57060	g1575061	Micro:143
capI	0.19	334	capsid protein	38.90	0.33	U10927	g506705	JBa 176:7005
lukM	0.22	308	LukM component of leukocidin and gamma-hemolysin:	45.09	0.32	D83951	g1230554	Unpublished
cap8N	0.24	295	capsular polysacc. biosynthesis	44.43	0.33	U73374	g1657653	JBa 179:1614
tagX	0.23	279	teichoic acid biosynthesis	44.23	0.31	U91741	g1913906	Unpublished
menc	0.26	333	o-succinylbenzoic acid (OSB) synthetase	45.31	0.30	U51132	g1255260	Unpublished
cap8M	0.21	185	capsular polysacc. biosynthesis	44.31	0.32	U73374	g1657652	JBa 179:1614
cap8G	0.24	374	capsular polysacc. biosynthesis	46.96	0.31	U73374	g1657646	JBa 179:1614
entE	0.21	257	entertoxin type E	40.82	0.33	M21319	g153002	JBa 171:4799
agrA	0.23	238	transducer	45.98	0.33	X52543	g46511	MGG 248:446
mts9	0.18	430	Sau96I DNA methyltransferase	38.36	0.34	X53096	g581567	NAR 18:4659
lukF-PV	0.23	322	LukF-PV like component	45.58	0.35	D83951	g1262748	Unpublished
plc	0.23	311	phosphatidylinositol-specific phospholipase C	42.32	0.31	L19298	g425478	IIm 61:5078
cap5N	0.26	295	Cap5N Capsid protein	45.86	0.31	U81973	g1773353	Unpublished
secE	0.23	60	tail-anchored membrane protein	48.72	0.42	U96619	g2078376	Unpublished
entC3	0.21	266	enterotoxin C3	41.82	0.33	X51661	g46571	MGG 220:329
glnR	0.23	122	glutamine synthetase repressor	46.59	0.32	X76490	g468509	JBa 176:1460
mene	0.25	492	o-succinylbenzoic acid (OSB) CoA ligase	45.92	0.28	U51132	g1255259	Unpublished
cap5I	0.17	369	Cap5I Capsid protein	42.01	0.33	U81973	g1773348	Unpublished
tsst-1	0.20	234	toxic shock syndrome toxin-1	37.88	0.39	J02615	g153123	JBC 261:15783
cap8L	0.31	401	capsular polysacc. biosynthesis	49.06	0.25	U73374	g1657651	JBa 179:1614
sau3AIR	0.19	489	restriction enzyme	40.72	0.36	A17958	g580669	Pat:EP0460673

Table 4-2 ***S. aureus*** **gene sequences (cont.)**

Gene	GC$_{3s}$	L	Gene description	ENc	Fop	Acc.#	Pid	Reference:
capB	0.20	228	capsular polysacc. biosynthesis	43.97	0.28	U10927	g506698	JBa 176:7005
capD	0.15	599	capsular polysacc. biosynthesis	39.26	0.35	U10927	g506700	JBa 176:7005
norA	0.22	388	fluoroquinolone resistance	44.54	0.36	D90119	g216975	JBa 176:4779
entB	0.22	266	enterotoxin B	44.66	0.34	M11118	g153000	JBa 166:29
qacA	0.13	188	antiseptic resistance	36.33	0.40	U22531	g46660	MM:2051+V84
sau3AIM	0.17	412	restriction enzyme	40.65	0.31	A17959	g809610	Pat:460673A
capK	0.13	449	capsular polysacc. biosynthesis	37.35	0.36	U10927	g506707	JBa 176:7005
mecI	0.17	123	methicillin resistance	45.88	0.41	X63598	g46615	FEBS 298:133
blaI	0.18	126	beta-lactamase repressor	47.33	0.40	M92376	g152970	AAC 36:2265
ORF	0.16	261	Sau96I restriction endonuclease	37.80	0.31	X53096	g46618	NAR 18:4659
entD	0.18	258	enterotoxin A precursor	44.53	0.35	M28521	g758691	JBa 171:4799
spsA	0.23	174	type-I signal peptidase SpsA	52.14	0.26	U65000	g1595809	JBa 178: 5712
capM	0.13	380	capsular polysacc. biosynthesis	34.37	0.38	U10927	g506709	JBa 176:7005
mecR1	0.18	585	methicillin resistance	44.46	0.34	X63598	g581566	FEBS 298:133
lgt	0.18	279	prolipoprotein diacylglyceryl transferase	38.21	0.33	U35773	g1016770	JBa 177:6820
ble	0.30	132	bleomycins	54.73	0.21	A31894	g1567208	PAT:W 9202230
llm	0.21	351	lipophilic protein	40.85	0.30	D21131	g2160281	JBa 176:4993
tetM	0.24	639	tetracycline resistance	47.28	0.25	M21136	g153115	AAC 34:2273
she	0.15	241	enterotoxin H	38.17	0.33	U11702	g510692	Unpublished
vatB	0.11	212	acetyltransferase	41.86	0.31	U19459	g1181627	AAC
agrB	0.21	423	signal transduction protein	44.31	0.34	X52543	g581546	MGG 248:446
sat	0.37	76	streptothricin acetyl transferase	46.18	0.24	U51473	g1272326	JMBT 6:218
capJ	0.17	391	capsular polysacc. biosynthesis	41.21	0.32	U10927	g506706	JBa 176:7005
cap8J	0.21	185	capsular polysacc. biosynthesis	36.68	0.32	U73374	g1657649	JBa 179:1614
capC	0.15	255	capsular polysacc. biosynthesis	39.04	0.31	U10927	g506699	JBa 176:7005
cap5H	0.24	208	putative O-acetyl transferase	55.61	0.22	U77308	g1673629	Unpublished
cap8H	0.20	359	capsular polysacc. biosynthesis	42.81	0.26	U73374	g1657647	JBa 179:1614
cat	0.11	219	chloramphenicol acetyltransferase	31.46	0.33	M58515	g152982	Unpublished
mraY	0.22	329	phospho-N-muramic acid-pentapeptide translocase	44.26	0.30	U94706	g2149894	Unpublished
capH	0.17	355	capsular polysacc. biosynthesis	39.89	0.34	U10927	g506704	JBa 176:7005
cap8K	0.20	412	capsular polysacc. biosynthesis	39.64	0.34	U73374	g1657650	JBa 179:1614
agrC	0.14	51	AgrC-31 truncated sensor protein	52.36	0.37	U85095	g1916239	Unpublished
lsp	0.15	163	prolipoprotein signal peptidase	40.31	0.28	M83994	g153045	FEBS 299:80
cap5J	0.21	388	Cap5J Capsid protein	46.31	0.28	U81973	g1773349	Unpublished
cap8I	0.21	464	capsular polysacc. biosynthesis	43.31	0.27	U73374	g1657648	JBa 179:1614
capG	0.17	172	capsular polysacc. biosynthesis	37.85	0.24	U10927	g506703	JBa 176:7005
vat	0.25	219	inactivates acetyltransferase	45.11	0.19	L07778	g398085	Gene 130: 91
capE	0.16	440	capsular polysacc. biosynthesis	38.46	0.26	U10927	g567036	JBa 176:7005
aphA-3	0.51	264	3'5'-aminoglycoside phosphotransferase	52.34	0.12	U51474	g1272327	JMBT 6:219
cap5K	0.24	401	capsular polysacc. biosynthesis	48.34	0.22	U81973	g1773350	Unpublished
capF	0.16	396	capsular polysacc. biosynthesis	39.52	0.29	U10927	g506702	JBa 176:7005

Table 4-2 ***S.*** *aureus* gene sequences (cont.)

Genes are listed in the order of their position on correspondence analysis axis 1. Gene is the gene name of each sequence. Where a gene has not been identified but has homology with another sequence it is labelled as an ORF, where it was not possible to identify a homologous sequence it is labelled as a unidentified reading frame URF. L is the length of the gene in codons. GC_{3s} is the G+C content at silent positions. EN_c is the effective number of codons used in a gene. F_{op} is the frequency of optimal codons. Acc. # is the GenBank/EMBL/DDBJ accession number. PID is the GenBank/EMBL/DDBJ protein identification number. References abbreviations are as in Table 3-1 with the following additions, Pat:Patent office; AAC:Antimicrob. Agents & Chemo; CJM:Cand. J. Microbiol; ZMEI: Zh. Mickrobiol. Epidem. Immun.; BBRC: Biochem. Biophys. Res. Commun.; JMB: J. Mol. Biol.; NAR: Nuc. Acid Res.; IMM:Infect. Immuno.; AMic: Arch. Microbiol.; JMB: J. Micro. Biotech.; PNAS: Proc. Nat. Acad. Sci.; CMBR:Cell. Mol. Biol. Res.; JBio:J. Biochem.

		High		Low				High		Low	
		RSCU	N	RSCU	N			RSCU	N	RSCU	N
Phe	UUU	0.81	(27)	1.70	(198)	Ser	UCU	1.41	(28)	1.04	(3)
	UUC	1.19	(40)	0.30	(35)		UCC	0.00	(0)	0.28	(9)
Leu	UUA	4.44	(128)	3.03	(182)		UCA	2.77	(55)	1.58	(50)
	UUG	0.28	(8)	0.83	(50)		UCG	0.00	(0)	0.44	(14)
	CUU	0.83	(24)	0.83	(50)	Pro	CCU	1.58	(39)	1.64	(34)
	CUC	0.03	(1)	0.17	(10)		CCC	0.08	(2)	0.53	(11)
	CUA	0.42	(12)	0.63	(38)		CCA	2.14	(53)	1.25	(26)
	CUG	0.00	(0)	0.50	(30)		CCG	0.20	(5)	0.58	(12)
Ile	AUU	1.60	(74)	1.50	(204)	Thr	ACU	2.00	(65)	1.21	(44)
	AUC	1.25	(58)	0.32	(44)		ACC	0.12	(4)	0.36	(13)
	AUA	0.15	(7)	1.17	(159)		ACA	1.75	(57)	1.77	(64)
Met	AUG	1.00	(41)	1.00	(68)		ACG	0.12	(4)	0.66	(24)
Val	GUU	1.80	(83)	1.22	(59)	Ala	GCU	2.18	(111)	1.57	(47)
	GUC	0.04	(2)	0.37	(18)		GCC	0.06	(3)	0.53	(16)
	GUA	1.93	(89)	1.62	(78)		GCA	1.55	(79)	1.50	(45)
	GUG	0.22	(10)	0.79	(38)		GCG	0.22	(11)	0.40	(12)
Tyr	UAU	1.32	(25)	1.62	(140)	Cys	UGU	1.11	(5)	1.18	(13)
	UAC	0.68	(13)	0.38	(33)		UGC	0.89	(4)	0.82	(9)
TER	UAA	3.00	(9)	1.33	(4)	TER	UGA	0.00	(0)	1.00	(3)
	UAG	0.00	(0)	0.67	(2)	Trp	UGG	1.00	(4)	1.00	(25)
His	CAU	1.10	(16)	1.20	(18)	Arg	CGU	4.68	(64)	0.39	(5)
	CAC	0.90	(13)	0.80	(12)		CGC	0.51	(7)	0.31	(4)
Gln	CAA	1.98	(122)	1.81	(39)		CGA	0.00	(0)	0.47	(6)
	CAG	0.02	(1)	0.19	(4)		CGG	0.00	(0)	0.70	(9)
Asn	AAU	0.72	(54)	1.65	(141)	Ser	AGU	0.91	(18)	1.83	(58)
	AAC	1.28	(97)	0.35	(30)		AGC	0.91	(18)	0.82	(26)
Lys	AAA	1.84	(211)	1.45	(131)	Arg	AGA	0.80	(11)	2.73	(35)
	AAG	0.16	(18)	0.55	(50)		AGG	0.00	(0)	1.40	(1)
Asp	GAU	1.15	(85)	1.56	(75)	Gly	GGU	2.81	(146)	1.55	(71)
	GAC	0.85	(63)	0.44	(21)		GGC	0.56	(29)	0.39	(18)
Glu	GAA	1.84	(179)	1.44	(77)		GGA	0.60	(31)	1.46	(67)
	GAG	0.16	(16)	0.56	(30)		GGG	0.04	(2)	0.59	(27)

Table 4-3 Putative optimal and rare codons of *S. aureus*. The high bias and low bias sets of genes contain 9 genes or 2381 and 2916 codons respectively. Those codons that occur significantly more often (p<0.01) in the highly biased dataset relative to the lower biased dataset are putatively considered optimal, and are indicated in red. Codons significant only at p<0.05 are indicated in purple. Those codons that are rare or absent as defined by Sharp *et al.* 1990, i.e. RSCU < 0.10, are indicated in blue.

When the codon usage of *S. aureus* is compared to *L. lactis* there is one notable difference, the leucine codon UUA is optimal (RSCU=4.44) in *S. aureus* but is avoided (RSCU=0.17) in highly expressed *L. lactis* genes. In all other cases in both species where a codon is optimal in either species, it is used in both at a higher frequency in high bias genes compared with low bias genes. Conversely, where a codon is rare in either species its usage decreases with increasing codon bias.

The F_{op} of the *S. aureus* genes was calculated using the 15 optimal codons identified above, and is tabulated in Table 4-2. It is intriguing to note that among the genes with high F_{op} values are two clinically important genes encoding protein A and immunoglobulin A binding protein.

The second most important trend in the codon usage variation of *S. aureus* genes is significantly correlated (r=0.79, p<0.001) with GRAVY, an index of amino acid hydrophobicity. As already seen above in the analysis of *L. lactis*, this trend is apparently due to selection pressure on codons encoding hydrophobic amino acids. In Figure 4-2 a group of putative membrane proteins with high GRAVY scores are delineated by a dashed ellipse.

4.2 Streptococcus mutans

S. mutans are non-motile Gram-positive cocci, found in characteristic chains, with a G+C content in the range 36-38 %. Based on DNA homology the species was divided into four subspecies (*mutans*, *rattus, cricetus,* and *sobrinus*) these were later elevated to species. These species are phenotypically similar and the name *S. mutans* remains widely used for this group. Most strains have enhanced growth under anaerobic conditions and produce acid from lactose. *S. mutans* synthesises several types of extracellular polysaccharides from sucrose that are important in the colonisation of tooth enamel. It was first isolated from carious human teeth, and is part of the natural flora of the mouth. It was largely overlooked for many years but a vast amount has been published since 1980, following the observation that many strains are highly cariogenic in experimental animals (Hamada and Slade 1980). It has also been isolated from cases of infective endocarditis (Sneath 1986).

4.2.1 Analysis of Codon Usage of *S. mutans*

The codon usage of *S. mutans* genes was investigated using the protocol described above, 171 annotated coding sequences were extracted from GenBank 95. The dataset was processed to remove partial sequences and horizontally transferred genes, as described above for *L. Lactis*, the *lac* operon genes were also removed from the dataset. This reduced the dataset to 57 genes, whose codon usage is tabulated in Table 4-1b.

Initial COA of the codon usage of these genes did not yield a principal axis that had a significant correlation with EN_c. Furthermore, the characteristic correlation between location on the principal axis and gene functional did not hold. A closer examination of the data revealed that the principal axis was correlated with GRAVY, and that the principal trend (explaining 15% of the total variation) was due to variation in amino acid composition. As the aim of the analysis was to attempt to identify codons used preferentially in highly expressed genes, it was necessary to remove the affect of amino acid composition bias by repeating the analysis with RSCU usage. The two principal axes from this second correspondence analysis explained 14.0% and 9.2% of the variation in RSCU respectively. When the sequences were sorted by their axis 1 ordination (see Table 4-4) ribosomal protein and metabolic enzyme genes were located towards one side of the distribution with regulatory and membrane proteins at the other. The principal axis was also significantly correlated ($r=0.68$, $p<0.001$) with EN_c. Despite information about amino acid composition being absent, the second axis remained correlated ($r=0.44$, $p<0.01$) with the GRAVY score, this correlation is similar to those previously described in the analyses presented above. A projection of each gene on the first two COA axes is presented in Figure 4-3. As discussed previously, these observations are a characteristic of species where selection is choosing between codons for translational efficiency. The codon usage of two groups of 5 genes (8% of the dataset) was contrasted to investigate if this variation was related to selection for a set of optimal codons. A 2-way chi squared contingency test identified 12 codons used preferentially in those genes with high codon bias, these encoded 12 amino acids (see Table 4-5). This set of codons contained the 6 universally optimal codons. A further four codons encoding 3 additional amino acids (the Serine codons UCA $p<0.01$ and UCU $p<0.05$) are used more frequently in the set of genes

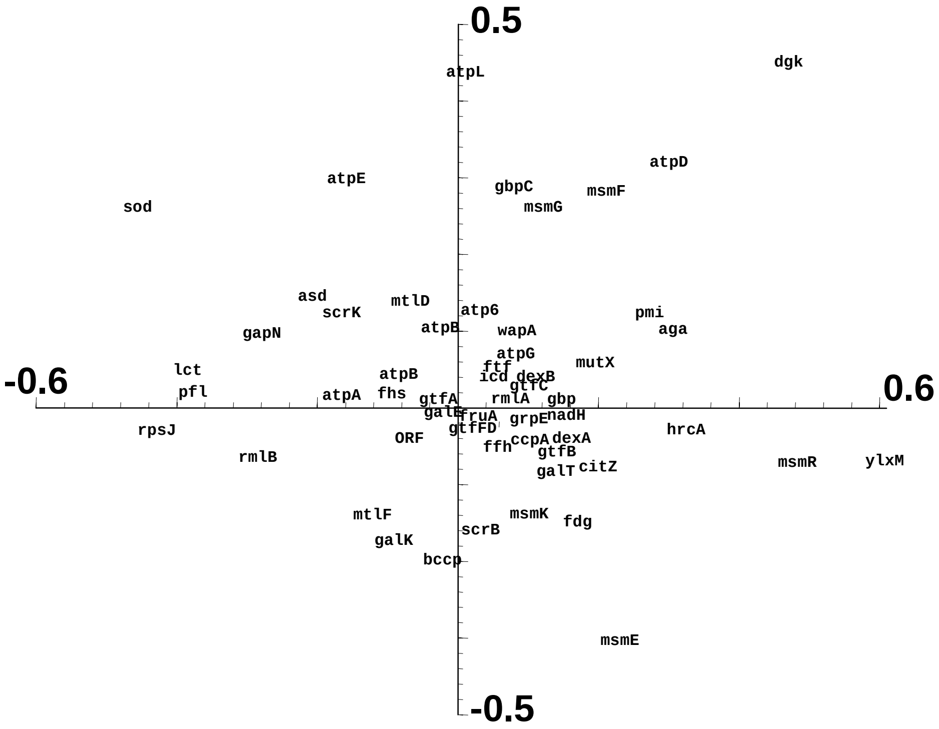

Figure 4-3 Correspondence analysis of the codon usage of 57 *S. mutans* genes. Each gene is plotted using its gene name at its coordinate on the first two axes produced by the analysis. The putatively highly expressed genes lie towards the left of axis 1**.**

Table 4-4 ***S. mutans*** gene sequences

Gene	GC_{3s}	L	Gene description	ENc	Fop	Acc.#	Pid	Reference:
sod	0.23	612	superoxide dismutase	33.76	0.33	S39782	g251295	JBa 174:4928
rpsJ	0.25	309	ribosomal protein	31.70	0.41	L29637	g467321	Unpublished
lct	0.22	987	lactate dehydrogenase	34.28	0.35	L42474	g833755	Unpublished
pfl	0.27	2328	pyruvate formate-lyase	37.71	0.39	D50491	g1129082	IIm 64:385
rmlB	0.38	1047	dTDP-glucose-4,6-dehydratase	45.79	0.36	D78187	g1813347	JBa 179:1126
gapN	0.22	1428	glyceraldehyde-3-phosphate dehydrogenase	37.13	0.32	L38521	g642667	JBa 177:2622
asd	0.20	1074	aspartate beta-semialdehyde dehydrogenase	35.79	0.26	J02667	g153562	JBC 262:3344
atpA	0.27	1506	ATPase, alpha subunit	42.58	0.26	U31174	g1773264	Gene 183:87
scrK	0.22	882	fructokinase	37.81	0.20	D13175	g287459	JBa 171:263
rmlC	0.29	597	dTDP-4-keto-L-rhamnose reductase	43.60	0.26	D78188	g1813346	JBa 179:1126
atpE	0.20	417	ATPase, epsilon subunit	41.95	0.22	U31177	g1773267	Gene 183:87
mtlF	0.25	438	mannitol-specific enzyme III	40.97	0.18	M94226	g153745	IIm 60:3369
galK	0.30	1173	galactokinase (EC 2.7.1.6).	45.87	0.24	U21942	g1877422	Gene 180:137
fhs	0.27	1671	formyl-tetrahydrofolate ligase; ATP-dependant synthetase	42.16	0.24	U39612	g1103865	JBa 179:1563
atpB	0.24	498	ATPase, b subunit	40.24	0.17	U31172	g1773262	Gene 183:87
mtlD	0.21	1149	mannitol-phosphate dehydrogenase	38.64	0.19	M94225	g153746	IIm 60:3369
ORF	0.20	1374	H2O-forming NADH Oxidase	39.02	0.24	D49951	g1199958	BBB 60:39
atpB	0.30	1410	ATPase, beta subunit	46.98	0.23	U31176	g1773266	Gene 183:87
bccp	0.23	393	biotin carboxyl carrier protein	48.47	0.25	M80523	g153584	BioT 14:209
galE	0.32	1002	UDP-galactose 4-epimerase	47.83	0.23	U21944	g1877424	Gene 180:137
gtfA	0.31	1464	sucrose phosphorylation	45.36	0.20	X08057	g47232	NAR 21:10398
atpL	0.22	204	ATPase, c subunit	38.60	0.22	U31170	g1773260	Gene 183:87
gtfFD	0.28	4293	glucosyltransferase-S enzyme	45.32	0.20	M29296	g153645	JGM 136:2099
fruA	0.28	4272	beta-fructosidase	46.10	0.20	L03358	g153634	IIm 60:4621
atp6	0.26	720	ATPase, a subunit	47.50	0.19	U31171	g1773261	Gene 183:87
scrB	0.23	1365	sucrose-6-phosphate hydrolase	45.11	0.20	X51507	g47259	IIm 56:1956
ffh	0.27	1551	signal recognition particle Ffh	42.85	0.20	U88582	g1850607	Unpublished
icd	0.28	1182	citrate synthase	43.39	0.20	U62800	g1421813	JBa 179:650
ftf	0.22	2394	fructosyltransferase	43.66	0.19	M18955	g153636	JBa 170:810
rmlA	0.27	870	glucose-1-phosphate thymidyltransferase	46.40	0.21	D78186	g1813345	JBa 179:1126
gbpC	0.21	1752	glucan-binding protein C	42.41	0.21	D85031	g1694933	IIm 65:668
atpG	0.26	879	ATPase, gamma subunit	45.90	0.19	U31175	g1773265	Gene 183:87
wapA	0.19	1338	wall-associated protein	42.72	0.20	M37842	g153858	MM 3:469
gtfC	0.32	4128	glucosyltransferase	48.72	0.20	M17362	g153641	JBa 169:4263
grpE	0.30	543	DNA girase	53.94	0.19	U78296	g2145132	Unpublished
msmK	0.33	1134	ATP-binding protein	48.12	0.21	M77356	g153741	IIm 56:1585
ccpA	0.27	1002	catabolite control protein A	46.86	0.16	AF001316	g2155300	Unpublished
dexB	0.27	1611	dextran glucosidase	48.19	0.15	M77352	g153742	IIm 56:1585
nadH	0.25	1533	NADH oxidase	44.61	0.20	D21803	g464216	Unpublished
msmG	0.29	834	membrane protein	43.81	0.23	M77355	g153739	IIm 56:1585
gbp	0.32	1692	glucan-binding protein	48.66	0.21	M30945	g153638	Unpublished
galT	0.36	1476	galactose-1-P-uridyl transferase	48.19	0.18	U21943	g1877423	Gene 180:137
gtfB	0.35	4428	glucosyltransferase	49.50	0.19	M17361	g153640	JBa 169:4263

Table 4-18 ***S. mutans*** gene sequences (cont.)

Gene	GC_{3s}	L	Gene description	ENc	Fop	Acc.#	Pid	Reference:
dexA	0.30	2553	dextranase	49.70	0.15	D49430	g1235734	MUK 141:2929
fdg	0.35	822	formamidopyrimidine-DNA glycosylase	54.96	0.13	D26071	d1005607	Unpublished
mutX	0.24	480	DNA repair	44.80	0.13	D78182	g1813348	JBa 179:1126
citZ	0.29	1119	citrate condensing enzyme	47.24	0.12	U62799	g1421813	JBa 179:650
msmF	0.28	873	membrane protein	44.51	0.21	M77354	g153738	IIm 56:1585
msmE	0.25	1263	sugar-binding protein	49.46	0.17	M77353	g153737	IIm 56:1585
scrR	0.29	963	regulator of scrB expression;	50.63	0.13	U46902	g1184967	Unpublished
pmi	0.26	951	mannosephosphate Isomerase	45.93	0.17	D16594	g451216	JGM 139:921
atpD	0.24	534	ATPase, delta subunit	42.25	0.15	U31173	g1773263	Gene 183:87
aga	0.27	2163	a-galactosidase	46.49	0.10	M77351	g153736	IIm 56:1585
hrcA	0.28	1038	repressor of class I heat shock gene expression	46.33	0.10	U78297	g2145131	Unpublished
dgk	0.26	414	diacylglycerol kinase	46.39	0.11	L12211	g409447	JBa 175:6220
msmR	0.28	837	regulatory protein	46.16	0.14	M77357	g455363	IIm 56:1585
ylxM	0.27	333	hypothetical 13.2 kDa protein	49.97	0.09	U88582	g1850606	Unpublished

Genes are listed in the order of their position on correspondence analysis axis 1. Gene is the gene name of each sequence. Where a gene has not been identified but has homology with another sequence it is labelled as an ORF, where it was not possible to identify a homologous sequence it is labelled as a unidentified reading frame URF. L is the length of the gene in codons. GC_{3s} is the G+C content at silent positions. EN_c is the effective number of codons used in a gene. F_{op} is the frequency of optimal codons. Acc. # is the GenBank/EMBL/DDBJ accession number. PID is the GenBank/EMBL/DDBJ protein identification number. References abbreviations are as in Table 3-1 and Table 4-2 with the following addition BioT:Biotechniques

		High		Low				High		Low	
		RSCU	N	RSCU	N			RSCU	N	RSCU	N
Phe	UUU	1.14	(40)	1.82	(82)	Ser	UCU	1.62	(23)	0.80	(15)
	UUC	0.86	(30)	0.18	(8)		UCC	0.21	(3)	0.69	(13)
Leu	UUA	0.55	(13)	1.71	(51)		UCA	3.25	(46)	1.43	(27)
	UUG	1.35	(32)	1.37	(41)		UCG	0.14	(2)	0.21	(4)
	CUU	3.04	(72)	1.61	(48)	Pro	CCU	0.98	(15)	2.13	(24)
	CUC	0.68	(16)	0.37	(11)		CCC	0.13	(2)	0.71	(8)
	CUA	0.13	(3)	0.47	(14)		CCA	2.56	(39)	0.89	(10)
	CUG	0.25	(6)	0.47	(14)		CCG	0.33	(5)	0.27	(3)
Ile	AUU	1.57	(58)	1.92	(78)	Thr	ACU	2.09	(59)	0.99	(20)
	AUC	1.35	(50)	0.47	(19)		ACC	0.39	(11)	1.28	(26)
	AUA	0.08	(3)	0.61	(25)		ACA	1.20	(34)	1.14	(23)
Met	AUG	1.00	(32)	1.00	(34)		ACG	0.32	(9)	0.59	(12)
Val	GUU	2.18	(68)	1.91	(45)	Ala	GCU	2.02	(85)	2.15	(56)
	GUC	0.61	(19)	0.85	(20)		GCC	0.62	(26)	0.65	(17)
	GUA	0.80	(25)	0.64	(15)		GCA	1.07	(45)	0.81	(21)
	GUG	0.42	(13)	0.60	(14)		GCG	0.29	(12)	0.38	(10)
Tyr	UAU	1.16	(46)	1.70	(63)	Cys	UGU	1.33	(6)	1.60	(4)
	UAC	0.84	(33)	0.30	(11)		UGC	0.67	(3)	0.40	(1)
TER	UAA	3.00	(5)	1.80	(3)	TER	UGA	0.00	(0)	0.00	(0)
	UAG	0.00	(0)	1.20	(2)	Trp	UGG	1.00	(22)	1.00	(13)
His	CAU	1.10	(28)	1.67	(36)	Arg	CGU	4.52	(55)	1.50	(17)
	CAC	0.90	(23)	0.33	(7)		CGC	1.23	(15)	0.88	(10)
Gln	CAA	1.76	(37)	1.41	(48)		CGA	0.08	(1)	0.97	(11)
	CAG	0.24	(5)	0.59	(20)		CGG	0.08	(1)	0.35	(4)
Asn	AAU	1.02	(49)	1.80	(82)	Ser	AGU	0.42	(6)	1.43	(27)
	AAC	0.98	(47)	0.20	(9)		AGC	0.35	(5)	1.43	(27)
Lys	AAA	1.58	(91)	1.29	(67)	Arg	AGA	0.08	(1)	2.12	(24)
	AAG	0.42	(24)	0.71	(37)		AGG	0.00	(0)	0.18	(2)
Asp	GAU	1.28	(76)	1.59	(81)	Gly	GGU	2.39	(70)	1.51	(28)
	GAC	0.72	(43)	0.41	(21)		GGC	0.99	(29)	0.86	(16)
Glu	GAA	1.94	(122)	1.44	(62)		GGA	0.58	(17)	1.03	(19)
	GAG	0.06	(4)	0.56	(24)		GGG	0.03	(1)	0.59	(11)

Table 4-5 Putative optimal and rare codons of *S. mutans* The high bias and low bias sets of genes contain 5 genes or 1761 and 1595 codons respectively. Those codons that occur significantly more often ($p<0.01$) in the highly biased dataset relative to the lower biased dataset are putatively considered optimal, and are indicated in red. Codons significant only a $p<0.05$ are indicated in purple. Those codons that are rare or absent as defined by Sharp *et al.* 1990, i.e. RSCU < 0.10, are indicated by blue.

with high codon bias but only at a significance level of $p<0.05$. One cause for the smaller number of optimal codons identified in *S. mutans* is the smaller number of genes in the *S. mutans* dataset compared with those used in the analyses of *S. aureus* and *L. lactis* codon usage. A further 7 codons could be classified as rare codons including GGG, AGG, CGG and CGA which were classified as rare in the analysis of *S. aureus* codon usage above, and *E. coli* and *B. subtilis* (see Table 4-6). No optimal codons were identified for the amino acids Val, Ala and Cys.

Overall these putative optimal and rare codons show a strikingly similarity to the codon usage of *B. subtilis*, *E. coli*, and *L. lactis*. There is strong support for the hypothesis that there is selection for a subset of preferred codons in *S. mutans* and that these codons can be considered as translationally optimal, however this selection appears to be weaker in *S. mutans* than in *S. aureus* and *L. lactis*. If the genes included in the analysis are a representative sample of *S. mutans* genes then there is only a very weak preference, if any, between codons for Val and Ala.

4.3 Comparison of the Codon Usage of *L. lactis, S. aureus* and *S. mutans*

L. lactis, *S. aureus* and *S. mutans* share a common preference for codons with A or U in the third codon position, with the GC_{3s} values for 90% of sequences falling into the range, 14-25% for *S. aureus*, 10-29% for *L. lactis*, and 18-31% for *S. mutans*. However, this preference does not extend to their choice of optimal codons, as several of the optimal codons have a C in the synonymous third position. All three species share a preference for the TAA termination codon in high bias (putatively highly expressed) genes. The optimal and rare codons of all three species show considerable similarity, with the exception of codon choice for leucine in *S. aureus* (see Table 4-21). As previously noted there is also a large overlap between the choice of optimal and rare codons in *E. coli, B. subtilis* with the codon choice in these three species (see Table 4-6).

Table 4-6 Comparison of optimal and non-optimal codons.

		L	E	B	S	S
		L	C	S	A	M
Phe	UUU					
	UUC	√	√	√	√	√
Leu	UUA		X		√	
	UUG	√				
	CUU	√		√		√
	CUC				X	
	CUA	X	X			
	CUG	X	√		X	
Ile	AUU					
	AUC	√	√	√	√	√
	AUA	X	X	X		X
Met	AUG					
Val	GUU	√	√	√	√	
	GUC					
	GUA			√		
	GUG	X				
Tyr	UAU					
	UAC	√	√	√		√
TER	UAA					
	UAG					
His	CAU					
	CAC	√	√			√
Gln	CAA	♦		√	√	♦
	CAG	X	√		X	
Asn	AAU					
	AAC	√	√	√	√	√
Lys	AAA	√	√	√	√	♦
	AAG	X				
Asp	GAU					
	GAC	√	√	√	√	♦
Glu	GAA	√	√	√	√	√
	GAG	X				X

		L	E	B	S	S
		L	C	S	A	M
Ser	UCU		√	√		♦
	UCC	X	√	X	X	
	UCA	√	√		√	√
	UCG	X		X	X	
Pro	CCU			√		
	CCC	X	X	X	X	
	CCA	√		√	√	√
	CCG		√			
Thr	ACU		√	√	√	√
	ACC	X	√	X		
	ACA					
	ACG	X				
Ala	GCU	√	√	√	√	
	GCC				X	
	GCA					
	GCG	X	√			
Cys	UGU					
	UGC		√			
TER	UGA					
Trp	UGG					
Arg	CGU	√	√	√	√	√
	CGC		√	√		
	CGA	X	X	X	X	X
	CGG	X	X	X	X	X
Ser	AGU					
	AGC		√			
Arg	AGA	X	X			X
	AGG	X	X	X	X	X
Gly	GGU	√	√	√	√	√
	GGC		√			
	GGA		X			
	GGG		X	X	X	X

Table 4-6. Comparison of optimal and non-optimal codons. Tabulation of codon choice for the species *L. lactis*, *B. subtilis*, *E. coli*, *Staphylococcus aureus,* and *Streptococcus mutans*. Legend key; LL=*L. lactis*; BS=*B. subtilis*; EC=*E. coli*; SA. *S. aureus*; SM=*S. mutans*. Optimal codons are indicated by red tick (√), those codon which occur significantly more often in high bias genes at p<0.05 are indicated by red diamonds (♦), non-optimal codons as defined by Sharp *et al*. 1990 are indicated by a blue cross (X). Termination codons and non-synonymous codons are shaded. Data for codon choice for *L. lactis*, *S. aureus* and *S. mutans* were taken from this thesis, *E. coli* and *B. subtilis* data were taken from Ikemura 1985 and Sharp *et al.* 1990.

Codon	S. mutans	S. aureus	L. lactis
UUA	0.55 (13)	4.44 (128)	0.17 (3)
UUG	1.35 (32)	0.28 (8)	2.20 (40)
CUU	3.04 (72)	0.83 (24)	3.19 (58)
CUC	0.68 (16)	0.03 (1)	0.44 (8)
CUA	0.13 (3)	0.42 (12)	0.00 (0)
CUG	0.25 (6)	0.00 (0)	0.00 (0)

Table 4-21 Differences in codon choice for leucine between *S. mutans*, *S. aureus* and *L. lactis* in high bias genes. Optimal codons are indicated in red and rare codons are indicated in blue.

5 The Lac Operon

5.1 The Lac Operon in Lactococcus lactis

Many *L. lactis* strains transport lactose across the cell wall by the phosphoenol-pyruvate dependent phosphotransferase system (PE-PTS). This lactose transport system has been found exclusively in the Gram-positive prokaryotes (de Vos and Vaughan 1994). In *L. lactis,* it is encoded as a single operon, which in some *L. lactis* strains is located chromosomally, but in the majority, the 8-kb *lac* operon is located on large plasmids. In the well characterised *lac* operon from *L. lactis* MG1820 the operon is flanked by a complete iso-IS1 element (de Vos *et al*. 1990).

The *lac* operon has been well defined and sequenced by several groups (de Vos and Vaughan 1994). In *L. lactis,* the *lac* operon contains nine genes; *lacR*, *lacA*-*lacG* and *lacX*. The gene products LacE and LacF transport lactose by the energetically efficient PE-PTS, during which lactose is phosphorylated. It is subsequently hydrolysed by LacG. The resulting product galactose 6-phosphate is degraded by the tagatose 6-phosphate pathway (LacAB-LacD). The *L. lactis lac* operon also includes the *lac* repressor *lacR* and an expressed open reading frame *lacX,* whose function is unknown (de Vos and Vaughan 1994). The organisation of the *L. lactis lac* operon is *lacRABCDFEGX* and Northern blot analysis has shown that the operon is transcribed as two partially overlapping polycistronic transcripts *lacABCDFE* and *lacABCDFEGX* (van Rooijen *et al*. 1991). The *lac* operon can be induced by lactose such that expression of transcripts increases up to 10 fold (de Vos *et al*. 1990; van Rooijen, van Schalkwijk and Devos 1991).

5.2 The Lac Operon in other Gram-positive bacteria

The PE-PTS is present in most commercial strains of Gram-positive lactic acid bacteria (de Vos and Vaughan 1994). Apart from *L. lactis* it has been best characterised in *Streptococcus mutans* (Honeyman and Curtiss 1993; Rosey and Stewart 1993), *Lactobacillus casei* (Chassy and Alpert 1989) and *Staphylococcus aureus* (Breidt and Stewart 1987; Oskouian and Stewart 1990; Rosey and Stewart 1989; Rosey *et al*. 1991). The PTS was actually first described in *Staphylococcus aureus* (Hengstenberg *et al*. 1967) in which it is chromosomally located

(Breidt and Stewart 1987). Chromosomally located PTS *lac* operons have also been reported in some strains of *Streptococcus mutans* (Honeyman and Curtiss 1993; Rosey and Stewart 1993).

The organisation and control of the *S. aureus* and *S. mutans* PE-PTS *lac* operons are similar to the (*lacRABCDFEGX*) structure of the *L. lactis lac* operon, with the exception that *lacX* appears to be exclusive to the *L. lactis lac* operon and the *lacR* repressor is translated in the opposite orientation. In *L. casei* the organisation (*lacTEGF*) and control of the *lac* operon is quite different (Chassy and Alpert 1989).

5.3 The Origin of the Lac Operon

Despite the importance of the PE-PTS lac operon (de Vos and Vaughan 1994), remarkably little is known about its origin. The operon is presumably quite young, on an evolutionary time-scale, as it is difficult to envisage a role of lactose metabolism prior to the radiation of the Eutheria. In conjunction with the conservation of structure and sequence of genes from the *lac* operon between *L. lactis*, *S. aureus* and *S. mutans* (see Table 5-1), this has been used to "strongly suggest that there is horizontal transfer of the *lac* genes between these Gram-positive bacteria" (de Vos *et al*. 1990). Attempts to establish a phylogenetic relationship for *lacE, lacG* and *lacF*, using the *L. casei* homologues as an outgroup, were inconclusive, the LacF supported one topology while the LacG sequences supported the alternative topology, the topology of LacE was poorly supported by bootstrap analysis (data not shown).

The similarity in the gene length (varying by less than 6 codons) and structure of the *lac* operon genes (excluding *lacR*) between *L. casei* and *S. aureus*, *S. mutans* and *L. lactis* (see Table 5-1) has been previously noted (de Vos *et al*. 1990). The lactose enzymes from *L. casei lac* are the most divergent in terms of sequence composition and retain between 43 and 54% identity. The LacR repressor has the lowest sequence identity (43-44%) of any of the gene pairs between *S. aureus*, *S. mutans* and *L. lactis* and this reflects the dissimilarities in orientation of this gene in these species. The remaining *lac* genes range from 61-63% identity for *lacC* to 82-90% for *lacG*.

	ll-sa	ll-sm	sa-sm	lc-ll	lc-sa	lc-sm
lacA	70.9	74.5	76.8			
lacB	85.5	76.2	79.6			
lacC	61.4	62.7	62.4			
lacD	73.4	78.2	72.4			
lacE	71.4	79.7	76.1	47.8	45.5	48.0
lacF	71.1	75.0	67.6	48.1	46.2	45.7
lacG	82.3	89.6	82.3	53.9	53.4	53.0
lacR	43.2	44.8	63.1			

Table 5-1 Pairwise sequence identity between the amino acid sequence of *lacA-G* and *lacR* genes between *L. lactis* (ll), *S. mutans* (sm), *S. aureus* (sa) and *L. casei* (lc). Values are uncorrected percentage sequence identity between each pair of proteins.

The overall codon preference of the three host species appeared to be broadly similar (see Table 3-1 and Table 4-1) sharing a preference in common for codons with A or T in the third codon position. If selection for translationally optimal codons has influenced codon choice in *lac* operon genes several conditions must be met;

- The operons must be infrequently horizontally transferred between species.
- There has been sufficient sequence divergence between the *lac* operons.
- There is selection to optimise translation efficiency in the hosts.
- The optimal translation of the *lac* operon genes improves the fitness of the host.

The total codon usage of the three *lac* operons are quite distinct from each other, particularly in the usage of the codons for the amino acids Leu, Val, and Ala, the codon usage of these amino acids is presented in Table 5-2. This suggests the operons may have been present in their current hosts for a considerable period of time. In Table 5-3 the synonymous substitution rates, as estimated using the method of Li *et al.* (1985) as amended by Li (1993), indicate that with the exception of *lacE* and *lacG* the silent sites are saturated with mutations. Therefore there has been ample opportunities for selection to choose between synonymous codons.

Putative optimal codons have been identified in all three host species, *L. lactis*, *S. aureus* and *S. mutans* which fulfils the third criteria. It has been previously demonstrated in this thesis that in *L. lactis*, the enzymes Tpi, Ldh, Pyk, and Pfk, which are involved in the metabolism of the products of the PE-PTS are highly expressed, see Table 3-2. This suggests the lac operon may be subject to selection of codon usage for optimal translation.

5.4 Adaptation of the Codon Usage of the Lac Operon to that of its Host

The first reported application of correspondence analysis to the investigation of codon usage by Grantham and co-workers in 1980 revealed the systematic variation of gene codon usage between genomes. Therefore to investigate how codon usage varies between the *L. lactis, S. aureus* and *L. lactis* genomic genes and whether the codon usage of the *lac* operons has

	S. aureus		*S. mutans*		*L. lactis*	
	RSCU (N)	Overall	RSCU (N)	Overall	RSCU (N)	Overall
Leu						
UUA	4.10 (112)	3.55	2.15 (57)	1.53	1.02 (28)	1.79
UUG	0.98 (26)	0.86	1.36 (36)	1.37	1.72 (47)	1.33
CUU	0.38 (10)	0.76	1.55 (41)	1.69	2.23 (61)	1.72
CUC	0.00 (0)	0.11	0.30 (8)	0.50	0.37 (10)	0.49
CUA	0.26 (7)	0.56	0.26 (7)	0.39	0.62 (17)	0.38
CUG	0.15 (4)	0.16	0.57 (15)	0.52	0.04 (1)	0.29
Val						
GUU	1.77 (55)	1.62	1.92 (59)	2.05	1.81 (56)	2.11
GUC	0.29 (9)	0.36	0.91 (28)	0.79	0.81 (25)	0.68
GUA	1.45 (45)	1.45	0.59 (18)	0.64	1.00 (31)	0.74
GUG	0.48 (15)	0.57	0.59 (18)	0.52	0.39 (12)	0.47
Ala						
GCU	1.10 (41)	1.36	2.02 (84)	2.00	1.77 (66)	1.69
GCC	0.27 (10)	0.25	0.55 (23)	0.63	0.51 (19)	0.57
GCA	2.04 (76)	1.87	1.20 (50)	1.05	1.42 (53)	1.36
GCG	0.59 (22)	0.52	0.22 (9)	0.32	0.30 (11)	0.39

Table 5-2 The codon usage for amino acids Leu, Val, Ala, and Gly in *lac* operon of *S. aureus*, *S. mutans*, and *L. lactis*. The total "overall" codon usage of each of the original datasets for *S. aureus*, *S. mutans*, and *L. lactis* are included for comparison.

a) K_s Synonymous substitution rates

	ll-sa	ll-sm	sa-sm	lc-ll	lc-sa	lc-sm
lacA	NA	123.1	155.6			
lacB	151.7	173.0	148.0			
lacC	96.2	71.1	92.7			
lacD	162.1	130.5	189.9			
lacE	44.1	37.9	42.0	59.6	78.1	64.0
lacF	165.9	120.6	NA	NA	NA	NA
lacG	43.6	45.6	54.5	110.2	106.1	94.6
lacR	220.9	NA	NA			

b) K_a Non-synonymous substitution rates

	ll-sa	ll-sm	sa-sm	lc-ll	lc-sa	lc-sm
lacA	22.1	18.1	14.8			
lacB	12.7	17.8	14.8			
lacC	41.2	39.0	40.4			
lacD	21.2	16.4	20.2			
lacE	38.0	34.3	31.6	69.0	72.1	68.6
lacF	19.8	27.0	15.2	43.9	44.6	49.4
lacG	20.5	18.3	20.5	51.5	48.7	47.1
lacR	53.5	56.0	27.4			

Table 5-3 Synonymous (K_s) and non-synonymous (K_a) substitution rates calculated for each pair of *lacA-G* and *lacR* genes from *L. lactis* (ll), *S. mutans* (sm), *S. aureus* (sa) and *L. casei* (lc). Values are the estimated number of substitutions per 100 nucleotides, calculated using the method of *Li et al* 1993 (see Methods). (NA = value cannot be calculated)

adapted towards the codon usage of its hosts, the RSCU of the genomic and *lac* operon genes were analysed using by COA. RSCU was used instead of codon usage because of the influence of amino acid composition on the COA of *S. mutans* sequences. The genes were plotted at their co-ordinates on the two principal axes generated by the analysis (see Figure 5-1).

The plot in Figure 5-1 allows the visualisation of the relationship between the RSCU of these genes. The genomic genes form three clusters but with some degree of overlap between them. The most widely separated clusters are those of *L. lactis* and *S. mutans*. This confirms the previous observation of a distinctive choice of codons for each species for particular amino acids but with a degree of overlap in codon usage. This plot also allows the direct comparison of the RSCU of the *lac* operon genes with genomic genes. The *lac* operon genes display a greater similarity in codon usage with the genomic genes from their hosts, than they do with each other. The clusters of *S. aureus* and *S. mutans* genomic genes overlay, and there is a corresponding overlap of their *lac* operon genes. The *L. lactis* genomic genes form a more distinct cluster and its *lac* operon genes are clearly located within this cluster. This reinforces the earlier observation that the codon usage of the operon genes are more similar to their host than to each other. This reflects that to some extent the codon usage of the operon has reached, or is reaching, an equilibrium with the over-all codon usage of its host. This also implies that these three operons have been present in these hosts (or hosts with similar codon usage) for a considerable period, an observation which is supported by the integration of these operons in the genomes of *S. aureus*, *S. mutans* and some *L. lactis* species.

5.5 Adaptation of the Codon Usage of the Lac Operon to the Optimal Codon Usage of its Host

The two-fold range of F_{op} values for *S. aureus* (F_{op} 0.30-0.58) and *L. lactis* (F_{op} 0.27-0.52) *lac* operon genes, presented in Table 5-4, suggests that the codon usage of at least some, if not all, *lac* operon genes from these species have been under selection for some form of optimal translation. It also suggests that this selection has not been equally effective at changing the codon usage of each gene.

Species	Gene	GC_{3s}	Length	ENc	F_{op}
ll	lacA	0.21	141	46.22	0.34
sa	lacA	0.20	142	39.53	0.49
sm	lacA	0.18	142	41.26	0.18
ll	lacB	0.28	171	38.52	0.44
sa	lacB	0.19	171	35.08	0.48
sm	lacB	0.18	171	47.26	0.18
ll	lacC	0.21	310	44.13	0.34
sa	lacC	0.21	310	40.48	0.38
sm	lacC	0.15	310	46.15	0.15
ll	lacD	0.23	326	38.07	0.52
sa	lacD	0.16	326	34.47	0.58
sm	lacD	0.21	325	46.52	0.21
ll	lacE	0.31	568	39.71	0.43
sa	lacE	0.22	572	37.89	0.39
sm	lacE	0.23	568	41.82	0.23
ll	lacF	0.29	105	49.91	0.27
sa	lacF	0.17	103	31.74	0.31
sm	lacF	0.15	104	58.57	0.15
ll	lacG	0.26	468	40.66	0.41
sa	lacG	0.19	470	35.75	0.43
sm	lacG	0.22	468	46.75	0.22
ll	lacR	0.24	261	51.91	0.21
sa	lacR	0.24	251	45.09	0.32
sm	lacR	0.18	251	49.04	0.18

Table 5-4 Comparison of the codon usage of the *lac* operon gene in *L. lactis*, *S. aureus,* and *S. mutans*.

Key to species, ll*: L. lactis*; sa*: S. aureus*; sm: *S. mutans*.

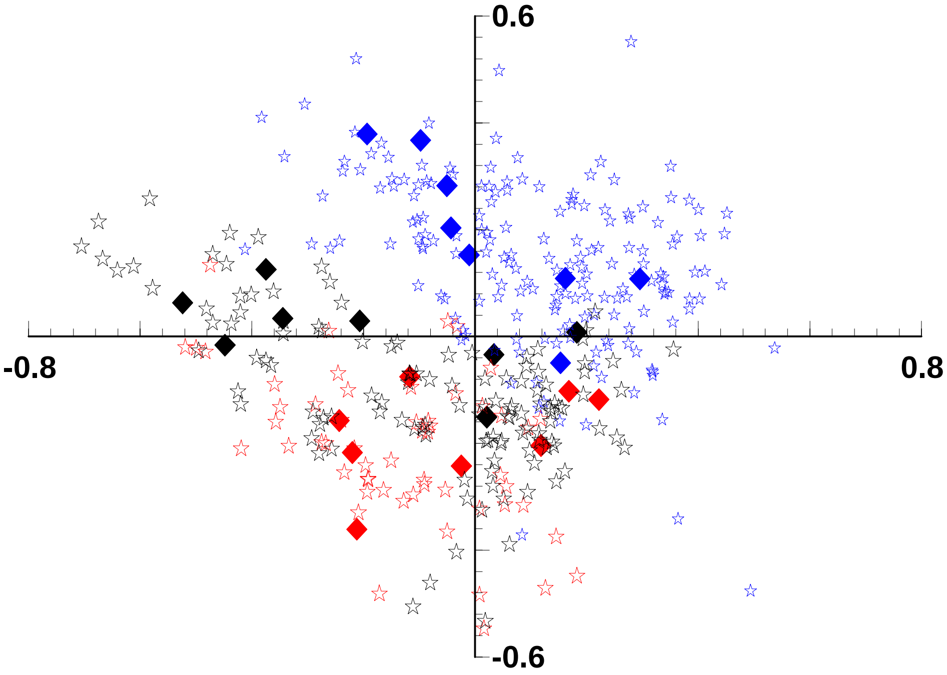

Figure 5-1 Correspondence analysis of RSCU of 383 *L. lactis*, *S. mutans*, and *S. mutans* genomic and *lac* operon genes. Genomic genes are represented by the hollow stars and *lac* operon genes by a solid diamond, *L. lactis* genes are coloured blue, *S. aureus* are black and *S. mutans* are red**.**

The F_{op} of genes from the *S. mutans lac* operon is lower and falls within a narrower range (i.e. 0.15-0.23) than for the *S. aureus* and *L. lactis* operon. This implies that selection has been less effective in *S. mutans* at adapting the codon usage towards the subset of optimal codons. The overall codon bias, as measured by EN_c, of six *S. aureus* and *L. lactis lac* genes (i.e. not *lacA* and *lacR*) was greater than for their *S. mutans* homologues despite *S. mutans* having the lowest GC_{3s} for 5 of the 8 genes (see Table 5-4).

While this comparison of F_{op} between species is interesting it must be emphasised that F_{op} was originally developed to quantify the adaptation of genes' codon usage within the same, or closely related species towards a set optimal codons. The comparison of F_{op} values and codon usage indices in general, between species is complex. This is especially true where there are an unequal number of optimal codons, or where the bias towards the subset of optimal codons is more pronounced in one of the species being examined.

COA has already been used to investigate the codon usage of genomic genes and to aid in the identification of optimal codons, and it was used to compare the codon usage of each host with its *lac* operon. However, if the codon usage of the *lac* operon and original genomic genes are simply combined and then analysed using correspondence analysis, the factors identified will not be the same as those found during the original analyses. Thus while the codon usage of both sets of genes could be compared with each other they wouldn't necessarily be compared in terms of the factor which was previously identified as being correlated with optimal codon usage. Fortunately, there is an alternative correspondence analysis methodology for directly comparing the codon usage of different datasets of genes. This is to use the vectors calculated during the correspondence analysis of one dataset to transform the codon usage of a different set of genes onto the same hyperspace as the original. CodonW, which has the functionality to do this type of COA, was used to compare the codon usage of genomic and *lac* operon genes. In Figure 5-2 the codon usage of the *lac* operon genes of each species are projected using the vectors calculated during the original codon usage analysis

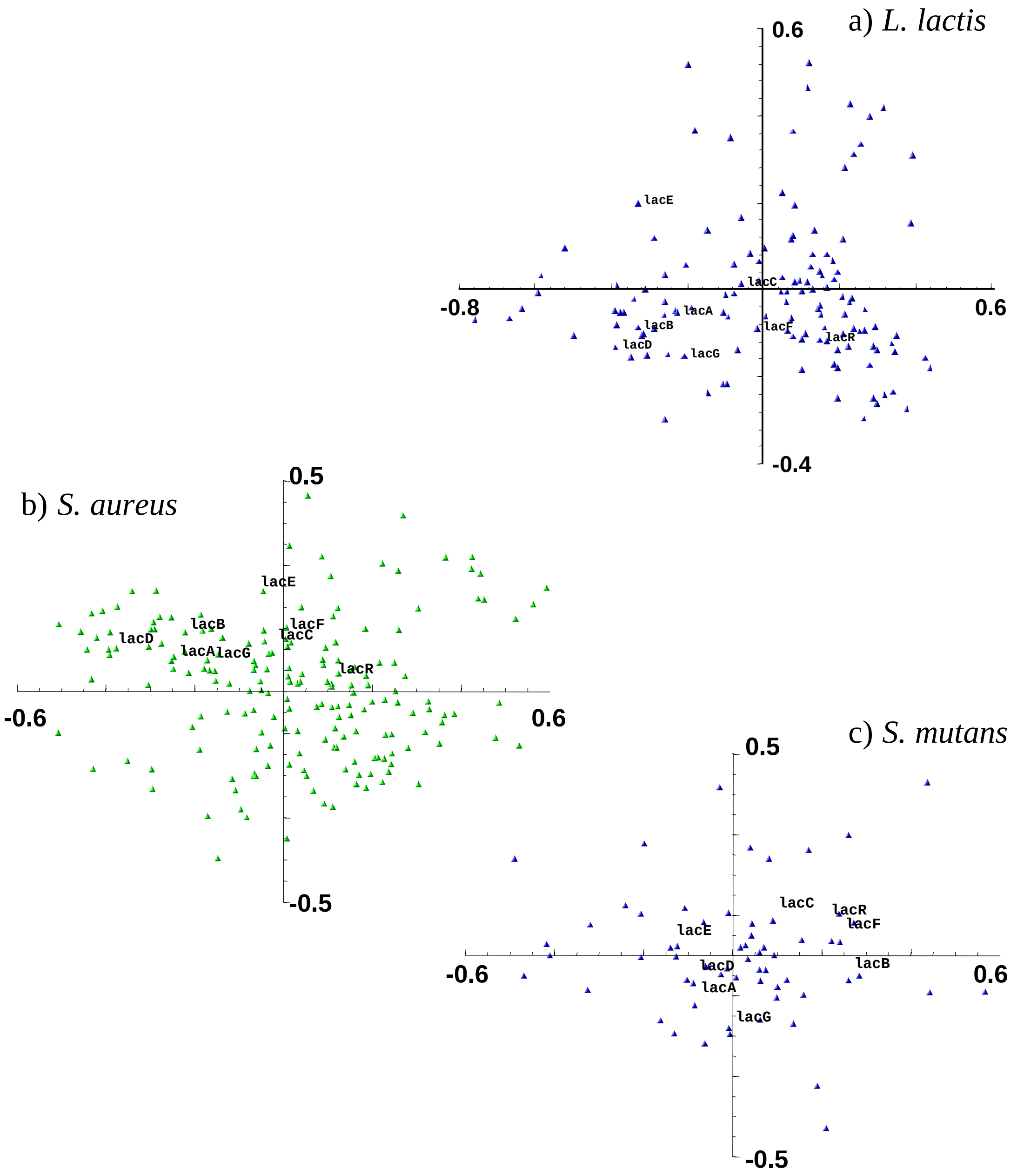

Figure 5-2 Projection of the *lac* operon genes of a) *L. lactis*, b) *S. aureus* and c) *S. mutans* using the correspondence analysis vectors for their respective hosts. The correspondence analysis vectors for each species were calculated from the codon usage of the genomic sequences that had been previously used to identify optimal codons in each species. The position of these genomic sequences on the first two axes is indicated by triangles.

for *L. lactis*, *S. aureus* and *S. mutans*, onto the two principal axes for each species. The original genes are also projected onto the same graph to facilitate the direct comparison of codon usage.

In Figure 5-2 the codon usage of *L. lactis* and *S. aureus lacD*, *lacB*, *lacA*, and *lacG* genes appears to have adapted, to differing degrees, towards the codon usage of highly expressed genes and are distributed along the principal trend. In both species the most translationally adapted *lac* gene is *lacD* (tagatose 1,6-diphosphate aldolase), which catalyses the final step of the tagatose 6-phosphate pathway. Surprisingly, the *lacD* termination codon is TAG, a termination codon that is rarely used in *L. lactis,* however a possible explanation is the presence of a second termination codon, TAA immediately following TAG, which may compensate for any leakiness of TAG itself. The only other *lac* operon gene to use TAG is the repressor protein *lacR* which also appears to have the least adapted codon usage. There is also some similarity in that the next two least adapted genes are *lacF* (soluble phosphocarrier-membrane transport) and *lacC* (tagatose 6-phosphate kinase). While the codon usage of *lacF* is perhaps unsurprising due to its functional role, the lack of adaptation of *lacC* is surprising. Other enzymes in the same pathway (*lacA, lacB*, and *lacD*) have adapted their codon usage. One possible explanation is that being a kinase it may have a high V_{max} and therefore high expression levels are unnecessary, it is also worth noting that the substrate of LacC is tagatose-6 phosphate which is the inducer for the lac operon (de Vos and Vaughan 1994). The similarity in the codon usage of *lacB* and *lacA* towards the optimal codon usage is reassuring as they form the dimer LacAB (galactose 6-phosphate isomerase). The principal difference between the codon usage of *L. lactis* and *S. aureus* operons (in terms of adaptation for optimal translation) is the membrane located *lacE* which appears much better adapted for optimal codon choice in *L. lactis* than *S. aureus*.

In contrast, the much weaker adaptation of *S. mutans lac* operon codon usage towards the codon usage of genes which have been putatively labelled as highly expressed is shown by the clustering of the *lacE*, *lacD*, *lacA*, and *lacG* around the origin. The codon usage of *lacC, lacF, lacB,* and *lacR* genes appearing similar to the codon usage of less highly expressed genes. This lack of adaptation of the codon usage of the *S. mutans lac* operon genes towards the subset of optimal codons could be caused by one or a combination of three factors: a) the operon not being present in *S. mutans* for a sufficient time to adapt its codon usage; b) it is not as highly expressed in *S. mutans* or c) it could be that the selection is not as effective to overcome mutational drift. The similarity between the overall codon usage of *S. mutans* and its *lac* operon does not support the first factor as a likely cause but it is very difficult to differentiate between the other two factors without detailed experimental results.

These observations tally well with the F_{op} indices presented in Table 5-4. One of the few discrepancies is in *L. lactis* where the F_{op} of *lacB* (0.38) which is lower than for *lacA* (0.48), the correspondence analysis places them close together. While adaptation for optimal codon usage, may explain the saturated K_s substitution rates for *lacD*, *lacB* and *lacA*, *lacG,* which has adapted codon usage for optimal translation, has a low K_s (43-54).

In both *S. aureus* and *L. lactis* there are differences, within their *lac* operon, in gene codon usage. This is presumably the result of selection for some form of translational efficiency. This suggests that the *lac* operon genes are transcribed at a range of expression levels. It is therefore an over simplification to compare overall *lac* operon codon usage with the codon usage of other operon, however the overall codon usage of the *lac* operon was used by Llanos *et al*. 1993, to infer that the *las* operon was highly expressed.

6 Cyanobacterium *Synechocystis* PCC6803

6.1 Introduction

Over 1500 cyanobacterial species are currently known - colonising a wide range of habitats. The cyanobacteria appeared about 2.7 billion years ago and are a diverse and widely distributed group of prokaryotes, phylogenetically distant but one of 11 major eubacterial phyla (Wilmotte 1994; Woese 1987). In fact they are more related to plant chloroplasts than to other eubacteria (Olsen *et al*. 1994). It is generally accepted that the progenitors of present day cyanobacteria were also the progenitors of plant plastids through endosymbiotic events (Kotani and Tabata 1998). Cyanobacteria (blue-green algae) are clearly distinguishable from other photosynthetic bacteria e.g. purple and green bacteria, in that they have the capacity to utilise H_2O as an electron donor.

Cyanobacteria have been used as model organisms for the study of oxygenic photosynthesis of higher plants where both chloroplast and nuclear genomes are involved. The cyanobacterium *Synechocystis* is also naturally transformable and therefore amenable to lab experimentation, some strains are capable of both photoautotrophic growth and in the absence of light and with the addition of glucose, heterotrophic growth (Pakrasi 1995).

6.2 Genome features

The genome of cyanobacterium Synechocystis PCC6803 has been completely sequenced (Kaneko *et al*. 1996). Total genome length was 3,573,470 bp with a G+C content of 47.7% (Kaneko *et al*. 1996). GeneMark (Borodovsky *et al*. 1994), was used to identify putative coding regions, the average gene length was 326 AA, the longest being 4199AA (Slr0408). ATG and GTG were the major initiation codons though TTG and ATG are also used (Kaneko *et al*. 1996). Gene density was average for bacterial genomes with 1 gene per 1.1kb. A total of 3168 potential protein encoding genes (representing 87% of the genome) were identified – 145 were identical to reported genes, 1257 were similar to reported genes and 340 to hypothetical genes, the remaining 1426 showed no database homology. 128 genes were related to oxygenic photosynthesis – photosystems I and II, phycobilisome formation, ATP synthesis, CO2 fixation and electron transport system (Kotani and Tabata 1998). Various elements of photosystem I and II commonly present in higher plants are absent in *Synechocystis* – *psa*G, *psa*H and *psa*N in system I, and *psb*P, *psb*Q, *psb*R, *psb*S, *psb*T and *psb*W for system II (Kaneko et al 1996).

A comparison of the deduced genes of *H. influenzae* with those of cyanobacterium *Synechocystis* PCC6803 reported that frequently several *Synechocystis* genes would have sequence identity with a single *H. influenzae* gene – suggesting gene multiplication in the *Synechocystis* genome (Kotani and Tabata 1998). Kotani and Tabata also noted more genes in common in the classes involved in basic cellular activities – DNA replication, restriction, modification, recombination, transcription, and translation, with fewer genes in common in classes such as photosynthesis (unsurprisingly!), respiration, and regulation.

Oligonucleotide analysis found an over-representation of homo-dinucleotides and high numbers of long homonucleotide runs and a very frequent palindrome GGCGATCGCC (high iterated palindrome -HIP1D). There are 2818 copies of HIP1D distributed randomly over the genome on average every 1268 bp, this distribution and a high density suggested that they contribute to genome-wide activities (Karlin *et al*. 1997). In protein-encoding regions 70% of these are translated as Ala-Ile-Ala. The origin and function of HIP1D in *Synechocystis* is unknown (Kotani and Tabata 1998). HIP1 is located at borders of gene deletions so may be a site for DNA recombination (Kotani and Tabata 1998). There is a related sequence in *H. influenzae*, the uptake signal sequence (USS), involved in DNA uptake but there is no evidence HIP1D is an uptake signal in *Synechocystis* (Kotani and Tabata 1998).

There are two structural RNA and tRNA operons, both have identical sequence, 870kb apart in reverse orientation and are 5028bp [16S rRNA]-[Ile-tRNA]-[23SrRNA]-[5S rRNA] (Kaneko et al 1995). Most algae and higher plant plastids also have 2 RNA gene clusters while eubacteria contain between 1 and 7 (Kotani and Tabata 1998). 42 putative tRNAs for 41 tRNA species were identified scattered throughout the genome, only trnI-GAU was duplicated. Most are transcribed as single units which is quite different from *E. coli* where 70% form clusters, interestingly the gene fMet-tRNA contains a 689 bp intron (Kaneko *et al*. 1996; Kotani and Tabata 1998).

6.2.1 Genes of possible exogenous origin

99 ORFs had sequence similarity to transposase genes – these could be classified into 6 groups and were spread over all of the genome, 26 of these at least appeared intact, remaining non-intact ORFs appear to be disrupted by frame-shift, deletion, and insertion of other IS like elements (Kaneko et al 1996). Each transposase is located within an IS like element (Kaneko et al 1996). The insertion sequences IS5S (871bp), IS4S (1299bp) and ISS1987 (949bp), are homologous to insertion sequences which occur in a wide range of hosts, and are believed to have been acquired by horizontal gene transfer (Cassierchauvat *et al*. 1997).

GeneMark has been shown to have difficulties identifying a subset of genes in *E. coli*, where the gene is believed to have been horizontally transferred (Borodovsky *et al*. 1995). In *Synechocystis*, those genes which could be identified by sequence homology but were not identified by GeneMark were putatively considered to have been horizontally transferred (Hirosawa *et al*. 1997). Examples include Sll0319 - β-lactamase precursor gene, ssl0562 – putative iron sulphur protein in photosystem I, slr0915 – endonuclease, and slr0317 – putative transposase gene of IS537.

6.3 Codon Usage

In Section 2.9.2.1 a correspondence analysis of the codon usage of the cyanobacterium *Synechococcus* sp. identified a set of putatively optimal codons, no evidence for selection for optimal codons had been previously identified in any cyanobacterial species. The availability of the whole genome sequence of the cyanobacterium *Synechocystis* PCC6803 provided the opportunity to investigate if the observation could be extended to another cyanobacterium. The codon usage of *Synechocystis* sp. PCC6803 had been reported prior to the publication of the whole genome, (Malakhov and Semenenko 1994; Schmidt *et al*. 1993) but no post-genome analysis has been published to date. Schmidt *et al.* (1993) surveyed the codon usage of 79 genes and compared this with the codon usage of the ribosomal proteins L1, L10, L11, L12, and L19, they also compared the codon usage of *aroC*. They concluded that the ribosomal proteins and *aroC* had the same pattern of codon usage as the global codon usage of *Synechocystis.* Malakhov and Semenenko (1994) surveyed the codon usage of 95 genes but concentrated on comparing it with the codon usage of dicots and monocots to investigate the feasibility of transgenic expression, but they concluded it was different to the codon usage of other organisms. They also noted that individual *Synechocystis* genes had similar codon usage to the entire dataset but they noted that the 95 genes that had been characterised probably belonged to a group of highly expressed genes.

6.3.1 Materials and Methods

A file containing 3168 potential coding sequences was obtained from the *Synechocystis* section of Cyanobase (URL http://www.kazusa.or.jp/cyano/cyano.html). All analyses were performed using CodonW. As the dataset was from a whole genome, duplicate sequences were not removed and there was no attempt to identify plasmid sequences. Sequences were annotated according to their descriptions in Table 1 of Kaneko *et al*. (1996).

6.3.2 Results

The total codon usage of the 3168 putative gene dataset is presented in Table 6-1a. Correspondence analysis of this dataset, excluding 23 ORFs which were shorter than 50 codons, yielded a first axis that explained 14.2% of the total variation in codon usage. However, the codon usage of the 99 transposases makes a large contribution to the variation represented by the first axis, as indicated by the location at the extreme sequences on axis 1 (see Figure 6-1). Their unusual codon usage combined with their association with IS elements makes it very probable that these transposases have been acquired by *Synechocystis* by horizontal gene transfer, therefore they were removed from the dataset.

This reduced the dataset to 3046 sequences, a correspondence analysis of the codon usage of this dataset yielded a first axis that explained 12.5% of the variation but which was highly correlated ($r=0.928$ $p<0.001$) with GC_{3s} (see Figure 6-2). Correlations between the principal trend and GC_{3s} have been previously observed when analysing the codon usage of *Mycoplasma genitalium*, where a systematic base composition variation around the genome is the principal source of codon usage variation

Table 6-1

a) Codon usage of 3168 sequences from cyanobacterium *Synechocystis* PCC6803. Total number of codons in table is 1036618.

		N	RSCU			N	RSCU			N	RSCU			N	RSCU
Phe	UUU	30490	1.47	Ser	UCU	8931	0.89	Tyr	UAU	17632	1.17	Cys	UGU	6448	1.24
	UUC	10983	0.53		UCC	16643	1.66		UAC	12483	0.83		UGC	3912	0.76
Leu	UUA	27022	1.38		UCA	4075	0.41	TER	UAA	1387	1.31	TER	UGA	627	0.59
	UUG	30611	1.56		UCG	4123	0.41		UAG	1154	1.09	Trp	UGG	16052	1.00
	CUU	10199	0.52	Pro	CCU	10148	0.76	His	CAU	11903	1.23	Arg	CGU	10579	1.22
	CUC	14562	0.74		CCC	26191	1.97		CAC	7420	0.77		CGC	12806	1.47
	CUA	14398	0.73		CCA	8167	0.61	Gln	CAA	35381	1.23		CGA	5380	0.62
	CUG	21113	1.07		CCG	8693	0.65		CAG	22038	0.77		CGG	14048	1.61
Ile	AUU	41766	1.93	Thr	ACU	14299	1.01	Asn	AAU	26153	1.25	Ser	AGU	15584	1.56
	AUC	18538	0.86		ACC	27547	1.94		AAC	15681	0.75		AGC	10717	1.07
	AUA	4652	0.21		ACA	6840	0.48	Lys	AAA	30461	1.41	Arg	AGA	4525	0.52
Met	AUG	20228	1.00		ACG	8203	0.58		AAG	12788	0.59		AGG	4899	0.56
Val	GUU	17078	0.98	Ala	GCU	20687	0.94	Asp	GAU	33344	1.28	Gly	GGU	20665	1.08
	GUC	11633	0.67		GCC	40038	1.83		GAC	18598	0.72		GGC	23794	1.25
	GUA	10885	0.62		GCA	10880	0.50	Glu	GAA	46370	1.48		GGA	13196	0.69
	GUG	30166	1.73		GCG	16133	0.74		GAG	16086	0.52		GGG	18585	0.98

b) Codon usage of 1648 sequences from *Synechocystis* PCC6803 which have some sequence identity with other known or predicted genes. Total number of codons in table is 618148.

		N	RSCU			N	RSCU			N	RSCU			N	RSCU
Phe	UUU	17037	1.44	Ser	UCU	4815	0.84	Tyr	UAU	9587	1.13	Cys	UGU	3853	1.24
	UUC	6557	0.56		UCC	10532	1.84		UAC	7445	0.87		UGC	2385	0.76
Leu	UUA	15233	1.31		UCA	1821	0.32	TER	UAA	741	1.37	TER	UGA	284	0.52
	UUG	18816	1.62		UCG	2409	0.42		UAG	601	1.11	Trp	UGG	8303	1.00
	CUU	5321	0.46	Pro	CCU	5511	0.71	His	CAU	7268	1.20	Arg	CGU	6372	1.21
	CUC	8945	0.77		CCC	15872	2.05		CAC	4814	0.80		CGC	8252	1.56
	CUA	8143	0.70		CCA	4339	0.56	Gln	CAA	20361	1.22		CGA	2975	0.56
	CUG	13061	1.13		CCG	5252	0.68		CAG	13100	0.78		CGG	9379	1.78
Ile	AUU	25738	1.97	Thr	ACU	7991	0.94	Asn	AAU	14030	1.19	Ser	AGU	8629	1.51
	AUC	11671	0.89		ACC	17482	2.06		AAC	9453	0.81		AGC	6096	1.07
	AUA	1825	0.14		ACA	3465	0.41	Lys	AAA	17962	1.40	Arg	AGA	1976	0.37
Met	AUG	12495	1.00		ACG	5044	0.59		AAG	7629	0.60		AGG	2718	0.51
Val	GUU	9850	0.91	Ala	GCU	12305	0.90	Asp	GAU	19734	1.26	Gly	GGU	12881	1.09
	GUC	7105	0.66		GCC	25923	1.89		GAC	11541	0.74		GGC	15164	1.29
	GUA	6487	0.60		GCA	6117	0.45	Glu	GAA	28416	1.50		GGA	7292	0.62
	GUG	19841	1.83		GCG	10568	0.77		GAG	9513	0.50		GGG	11823	1.00

N is codon frequency, RSCU is relative synonymous codon usage.

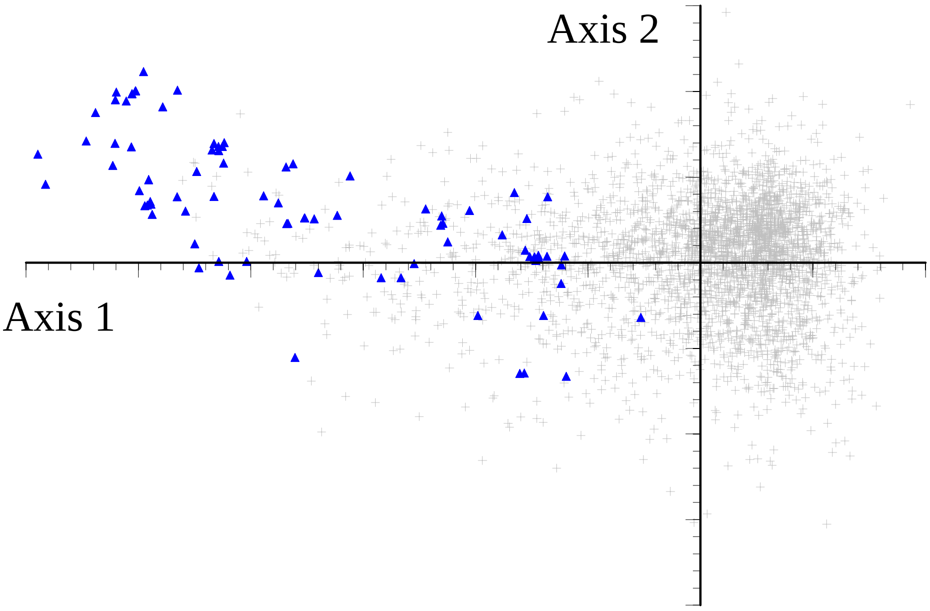

Figure 6-1 Projection onto the first two axes of a correspondence analysis of 3145 (3168 ORFs excluding those shorter than 50 codons in length) genes from *Synechocystis* PCC 6803, the location of transposases are represented by blue triangles.

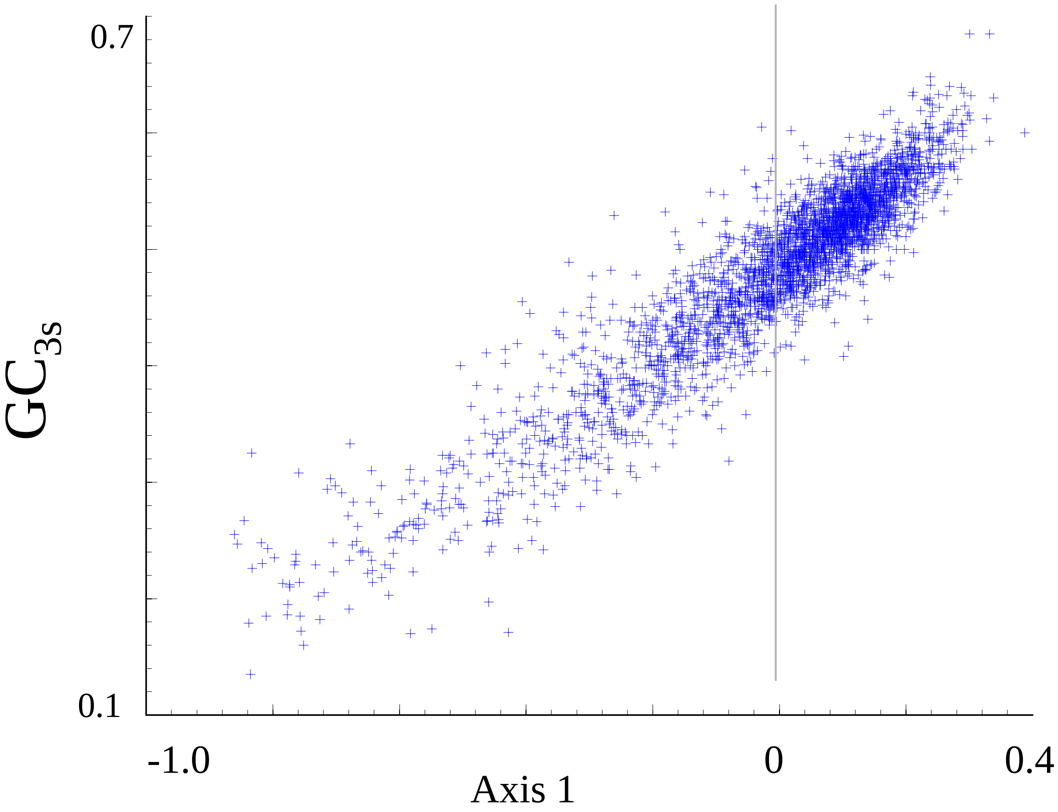

Figure 6-2 Correlation between GC$_{3s}$ and the Axis 1 ordination from a correspondence analysis of 3046 *Synechocystis* PCC6083 genes, the grey line marks the origin of axis 1.

(Kerr *et al*. 1997; McInerney 1997) and in *Borrelia burgdorferi* where the principal source for variation in codon usage is a strong strand-specific skew in nucleotide composition (McInerney 1998). However, strand compositional asymmetry is not observed in the *Synechocystis* sp. genome (Mrazek and Karlin 1998) and there is no evidence of systematic variation between gene location and GC_{3s} (data not shown). A projection of genes onto the first two correspondence analysis axis (see Figure 6-3) reveals that the majority of the sequences which have Axis 1 coordinates in the range –1 to -0.25 (i.e. those genes with the lowest GC_{3s}) have only been predicted by GeneMark and have no database hits to support their prediction. In fact over 40% (1426 ORFs) of the predicted protein coding regions of *Synechocystis* had no supporting database matches, it seems likely that a high proportion of these poorly supported genes are not real protein coding regions.

A correspondence analysis of the codon usage of 1626 *Synechocystis* putative coding regions, (i.e. all putative coding regions with supporting database matches but excluding the transposases and ORFs shorter than 50 codons in length), yielded a first axis that account for 9.9 % of the total variation in codon usage, and a second axis that accounted 6.8% of the variation. Axis 1 was significantly correlated with GC_{3s} ($r=0.81$, $p<0.001$) but was also significantly correlated with EN_c ($r=0.39$, $p<0.001$), axis 2 was correlated with aromaticity ($r=0.37$, $p<0.001$) and axis 3 was significantly correlated ($r=0.66$, $p<0.001$) with the GRAVY score (see Table 6-2). A projection of the genes, labelled according to their functional groups as assigned by Kaneko *et al.* 1996, onto the first two correspondence analysis axes allows the comparison of axis 1 position with gene function (see Figure 6-4). Ribosomal genes, genes involved in photosynthesis and heat shock protein genes are located towards the right-hand side of the distribution of genes, while amino-acid biosynthesis genes lie towards the centre and several regulatory proteins lie to the left of the distribution. Caution must be exercised when examining the ordination of the regulatory proteins as *Synechocystis* utilises two-component regulatory systems which are related to those of higher eukaryotes, therefore some of the genes labelled as regulatory proteins are actually protein kinases. The regulatory protein with the most positive axis 1 coordinate (0.29) is slr0947 (*ycf*27) which has sequence similarity to histidine protein kinases (Bartsevich and S.V. 1995). Another two regulatory genes slr0741 (*amiC*) and slr0447 (*phoU*) which also have positive axis 1 coordinates (0.23 and 0.20 respectively) have little sequence similarity

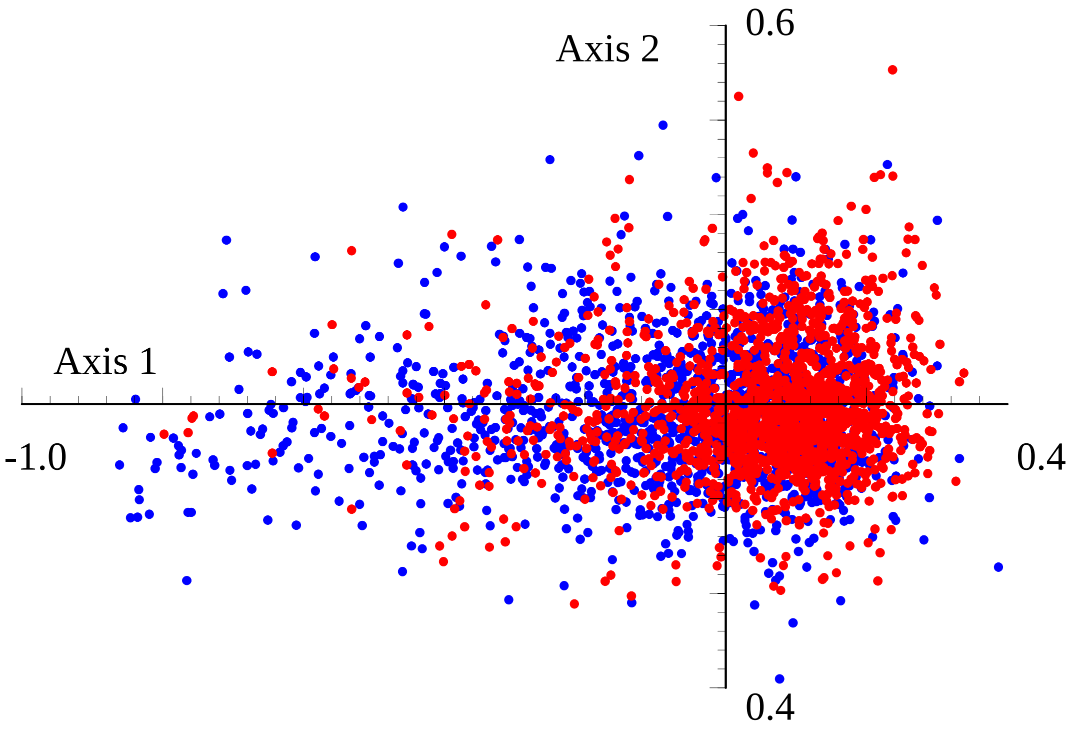

Figure 6-3 Projection onto the first two axes of a correspondence analysis of 3046 genes from *Synechocystis* PCC 6803. Genes which have only been predicted by GeneMark are shown in blue, genes which have similarity matches with sequences in the public databases are shown in red.

	Nc	GC3s	GC	Gravy	Aromo
Axis 1	-0.393	0.815	0.723	-0.030	-0.211
Axis 2	-0.109	-0.133	-0.214	0.328	0.371
Axis 3	0.103	0.213	0.264	0.660	0.080
Axis 4	0.092	-0.019	0.022	-0.024	-0.418

Table 6-2 Correlation between the codon and amino acid usage indices, EN_C, GC_{3s}, GC, GRAVY and Aromaticity, with the first four axes from a correspondence analysis of 1624 *Synechocystis* PCC 6803 genes**.**

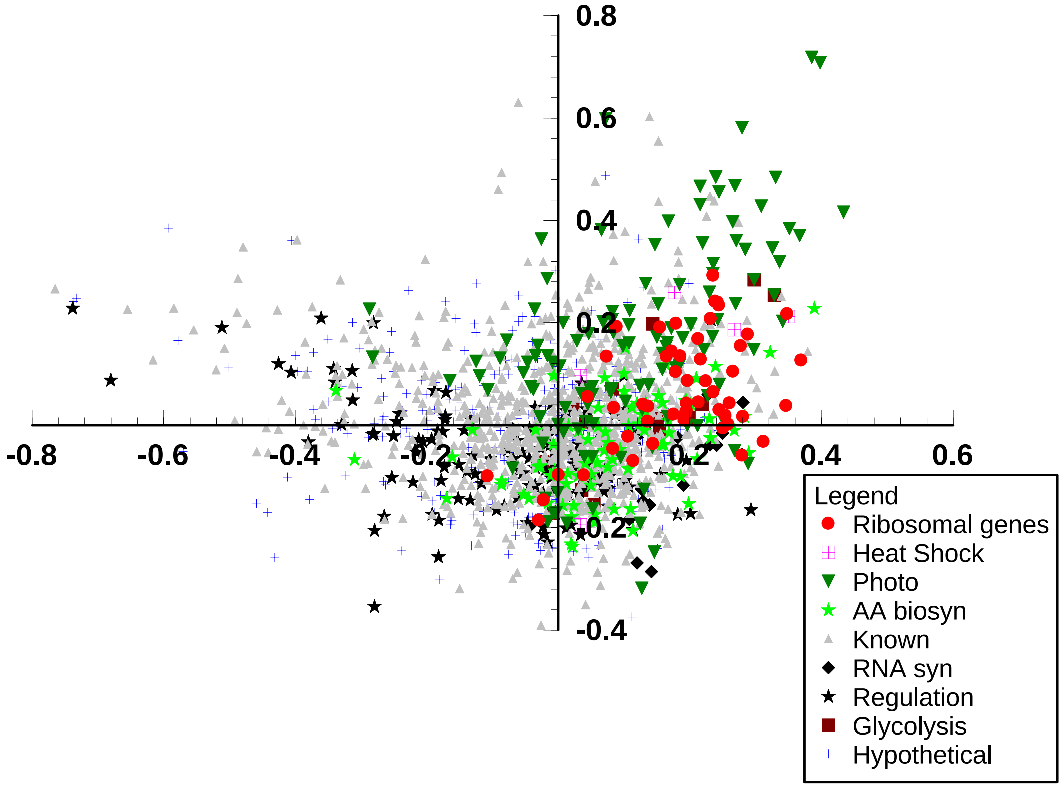

Figure 6-4 Projection of 1626 Synechocystis genes onto the first two axes of a correspondence analysis of the dataset. Genes have been labelled according to their function as assigned *by Kaneko et al.* 1996. Key to legend is: Ribosomal genes- Ribosomal protein genes; Heat shock – heat shock proteins; Phot- photosynthetic genes; AA biosyn- amino acid biosynthesis proteins; Known- other proteins with assigned function but which did not fall into one of the displayed categories; Regulation- regulation of cell, Glycolysis- respiration; Hypothetical- no function assigned; RNA syn- RNA synthesis.

(13.4% and 15% respectively) with the proteins which were used to identity them, and may have been misidentified.

Putative optimal codons were identified from this dataset using the methodology described previously, using 100 genes sampled from both extremes of the distribution along axis 1, and are presented in Table 6-3. The “universally” optimal codons, UUC, UAC, AUC, AAC GAC, and GGU are amongst the putative optimal codons. Putative optimal codons were identified for 17 synonymous codon groups, with the exception of Lys (AAG is significant at $p<0.05$), two putative optimal codons were identified for Gly, Arg, and Ser, and three were identified for Leu.

6.3.3 Comparison with the putative optimal codons of Synechococcus sp.

Synechococcus sp. and *Synechocystis* share 16 putative optimal codons (in addition to the universal codons these are CUC, CUG, CAC, GAA, CCC, ACC, GCU, CGC, AGC and GGC). *Synechocystis* has an extra putative optimal codon (UCC) for Ser, Leu (UUG) and Cys (UGC). The putative optimal codon for Gln has changed from CAA in *Synechococcus* to CAG in *Synechocystis*, and one of the putative optimal codons for Arg has changed from CGU to CGG.

Eight codons were classified as rare in the high bias dataset of *Synechococcus sp.* but only AUA is classified as rare in the *Synechocystis* dataset. The selective pressure, which results in the avoidance of rare codons, does not seem to be as strong in *Synechocystis*, though it must be noted that the *Synechococcus* dataset was much smaller. Each of the rare *Synechococcus* codons is however still used less often in the high bias *Synechocystis* dataset, with the exception of CUA where the difference is marginal

In *Synechocystis* sp. PCC6803 and *Synechococcus* sp. photosynthetic and ribosomal protein genes which might be expected to be highly expressed, have some of the strongest biases in codon choice. They preferentially use a subset of codons, including those which have been shown to be optimal in other eubacterial species. These observations are characteristic of

		High		Low				High		Low	
		RSCU	N	RSCU	N			RSCU	N	RSCU	N
Phe	UUU	0.88	(543)	1.65	(1536)	Ser	UCU	0.85	(248)	1.27	(561)
	UUC	1.12	(685)	0.35	(330)		UCC	2.73	(791)	1.01	(446)
Leu	UUA	0.71	(352)	2.07	(1574)		UCA	0.12	(34)	0.79	(350)
	UUG	1.70	(839)	1.25	(947)		UCG	0.29	(84)	0.38	(170)
	CUU	0.39	(195)	0.80	(605)	Pro	CCU	0.70	(253)	1.15	(492)
	CUC	1.20	(593)	0.50	(378)		CCC	2.61	(948)	1.32	(566)
	CUA	0.52	(257)	0.79	(601)		CCA	0.23	(83)	1.04	(444)
	CUG	1.48	(730)	0.60	(452)		CCG	0.47	(171)	0.49	(209)
Ile	AUU	1.69	(1216)	1.98	(2141)	Thr	ACU	0.65	(300)	1.44	(789)
	AUC	1.30	(934)	0.60	(648)		ACC	2.83	(1315)	1.21	(663)
	AUA	0.01	(10)	0.42	(459)		ACA	0.14	(63)	0.90	(492)
Met	AUG	1.00	(842)	1.00	(614)		ACG	0.39	(180)	0.44	(243)
Val	GUU	0.94	(607)	1.55	(922)	Ala	GCU	1.08	(821)	1.36	(930)
	GUC	0.71	(457)	0.62	(371)		GCC	2.13	(1623)	1.25	(851)
	GUA	0.71	(457)	0.85	(506)		GCA	0.22	(166)	0.84	(573)
	GUG	1.65	(1064)	0.98	(586)		GCG	0.58	(441)	0.56	(380)
Tyr	UAU	0.77	(362)	1.46	(954)	Cys	UGU	1.10	(151)	1.34	(286)
	UAC	1.23	(582)	0.54	(355)		UGC	0.90	(124)	0.66	(141)
TER	UAA	1.68	(56)	1.53	(51)	TER	UGA	0.15	(5)	0.45	(15)
	UAG	1.17	(39)	1.02	(34)	Trp	UGG	1.00	(346)	1.00	(614)
His	CAU	0.69	(219)	1.45	(525)	Arg	CGU	1.49	(475)	1.34	(382)
	CAC	1.31	(419)	0.55	(201)		CGC	1.76	(561)	0.98	(279)
Gln	CAA	1.18	(738)	1.39	(1537)		CGA	0.23	(72)	0.95	(271)
	CAG	0.82	(514)	0.61	(682)		CGG	2.20	(700)	0.92	(262)
Asn	AAU	0.70	(437)	1.42	(1488)	Ser	AGU	0.82	(238)	1.74	(770)
	AAC	1.30	(803)	0.58	(607)		AGC	1.19	(346)	0.81	(358)
Lys	AAA	1.44	(1267)	1.50	(1455)	Arg	AGA	0.15	(49)	1.00	(286)
	AAG	0.56	(487)	0.50	(479)		AGG	0.16	(51)	0.83	(236)
Asp	GAU	1.03	(982)	1.44	(1508)	Gly	GGU	1.66	(1163)	1.19	(763)
	GAC	0.97	(917)	0.56	(581)		GGC	1.33	(933)	1.01	(645)
Glu	GAA	1.60	(1896)	1.52	(1884)		GGA	0.33	(232)	1.01	(644)
	GAG	0.40	(478)	0.48	(599)		GGG	0.67	(469)	0.79	(502)

Table 6-3 Putative optimal and rare codons of *Synechocystis*. The high bias and low bias sets of genes contain 33413 and 40233 codons respectively. Those codons that occur significantly more often (p<0.01) in the highly biased dataset relative to the lower biased dataset are putatively considered optimal, and are indicated in red. The codons which are significant only at p<0.05 are marked in purple. Codons that are rare or absent as defined by Sharp *et al.* 1990, i.e. RSCU < 0.01, are indicated by blue.

species where selection for some form of translational efficiency can discriminate between synonymous codons. As the putative optimal codons fulfil the necessary criteria they can be tentatively considered as translationally optimal.

7 Conclusions

A primary objective of this project was to develop a portable, flexible and powerful codon usage analysis programme. CodonW was designed to fulfil these criteria and has been extensively tested within the scope of this research project, all codon usage analyses reported herein were performed using this single programme. The programme has also been released under the terms of the GNU public software licence and has been widely distributed. It has been installed by international genome sequencing centres where it is used to analyse the codon usage of newly sequenced bacterial genomes.

To demonstrate some of the functionality of CodonW the amino acid usage of *S. typhimurium* was analysed using correspondence analysis. This investigation confirmed, as for *E. coli*, that the most important trend in the variation of *S. typhimurium* amino acid usage is for the usage of hydrophobic amino-acids, and the third most important trend is aromaticity. However, in *S. typhimurium*, the correlation between expression level and amino acid usage (the second most important trend in *E. coli)* is much weaker than previously reported for *E. coli* and it is uncertain if translational selection is choosing between amino acids in this species. The codon usage of *S. cerevisiae* was surveyed using standard indices and in-built yeast codon preference tables. The distribution of codon usage indices was unimodal and positively skewed, it was estimated that only between 3% and 4% of yeast genes had high codon bias.

The codon usage of the genes from the whole *S. cerevisiae* genome were also analysed using the correspondence analysis option of CodonW. Genes were automatically assigned as being putatively highly and less highly expressed, these sequences were then used to identify optimal codons and to automatically calculate CAI fitness values. The putative optimal codons identified using the default settings for CodonW, which included 5% of the dataset in each category, identified all the yeast optimal codons that had been previously reported and an extra codon CGU. However, the calculated codon fitness values did not tally with those previously published. By restricting the number of genes automatically assigned to 50, the optimal codons matched those previously reported and CAI calculated from the new codon fitness values were very significantly correlated with CAI calculated with codon fitness values derived from experimentally verified highly expressed genes. This demonstrated it was feasible to automatically identify highly expressed genes, putative optimal codons and to assign codon fitness values, but the number of highly expressed genes is likely to be approximately 1%.

There have been no previous reports of selection for translational efficiency choosing between codons in cyanobacteria. The codon usage of three strains of *Synechococcus* were examined and strong evidence was found that there is selection for a preferred subset of codons and that this selection was for translational efficiency. This observation was extended by an analysis of the codon usage of the whole genome of *Synechocystis* sp. PCC6803, which was published during the period of this project.

Correspondence analysis of the whole genome clearly identified the transposases as having atypical codon usage. It also found that some of the putative protein-coding regions predicted by GeneMark but for which there was little supporting evidence had unusually low GC_{3s}. These putative protein-coding regions may have been the result of over-prediction by GeneMark. The correspondence analysis of predicted coding regions which had additional supporting evidence, found that in general ribosomal and photosynthetic genes had more biased codon usage and that compared with a group of genes with low codon bias they preferentially used a subset of 22 codons. However, the codon bias of the ribosomal and photosynthetic genes did not appear to be particularly strong. These codons were tentatively identified as optimal. This set of optimal codons overlapped with 16 of those identified in *Synechococcus* sp. The optimal codon for Gln is CAA in *Synechococcus* and CAG in *Synechocystis*, and one of the optimal codons for Arg has changed from CGU in *Synechococcus* to CGG in *Synechocystis*. There appears to be a greater number of rare codons in *Synechococcus.*

The codon usage of the low G+C gram positives *L. lactis*, *S. aureus* and *S. mutans* was surveyed for evidence of selection for translational efficiency. Evidence for selection for translational efficiency was found in all three species. In *L. lactis* 17 optimal codons were identified. There is evidence that the *L. lactis* lactic acid fermentation pathways are under selection, as codon usage of several of these genes had a higher frequency of optimal codons relative to their *B. subtilis* homologues. The presence of the “very unusual” *pab*B gene in the initial dataset clearly demonstrated that simple tabulation of codon usage can be fraught with difficulties.

In *S. aureus* 15 optimal and 11 rare codons were identified. In *S. mutans* the influence of selection for translational efficiency was initially masked by amino acid composition and appears weaker than in the other two species, only 12 optimal codons could be identified but the *S. mutans* dataset was also the smallest. The choice of *L. lactis*, *S. aureus* and *S. mutans* for rare and optimal codons (where they occurred) was broadly similar with the exception of the choice of optimal codon for Leu, where each species had a different preference.

The PEPTS-*lac* operon has only been reported in *L. lactis*, *S. aureus* and *S. mutans*. The codon usage of each set of operon genes was compared with the codon usage of its host. Overall, the codon usage of each operon is similar to that of its host implying that the operons may have been present in their current hosts for a considerable period of time. There is a higher frequency of optimal codons in the lac operon genes *lacD*, *lacC*, *lacB* and *lacA* of *L. lactis* and *S. aureus*. Correspondence analysis shows that the codon usage of these genes is similar to the codon usage of highly expressed genes in these species. Therefore, the codon usage of the *lac* operon appears to be under selection for translational efficiency. The variation in codon usage between the genes of the *L. lactis* and *S. aureus lac* operons strongly suggests that they are expressed at a range of levels. The codon usage of the *S. mutans lac* operon does not appear to be as adapted for translational efficiency.

8 Bibliography

Adamski, F. M., K. K. McCaughan, F. Jorgensen, C. G. Kurland and W. P. Tate, (1994). The concentration of polypeptide-chain release factor-1 and factor-2 at different growth-rates of *Escherichia coli*. Journal of Molecular Biology **238:** 302-308.

Airenne, K. J., P. Sarkkinen, E. L. Punnonen and M. S. Kulomaa, (1994). Production of recombinant avidin in *Escherichia coli*. Gene **144:** 75-80.

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

Akashi, H., (1995). Inferring weak selection from patterns of polymorphism and divergence at silent sites in *Drosophila* DNA. Genetics **139:** 1067-1076.

Alffsteinberger, C., (1984). Evidence for a coding pattern on the non-coding strand of the *E. coli* genome. Nucleic Acids Research **12:** 2235-2241.

Alffsteinberger, C., and R. Epstein, (1994). Codon preference in the terminal region of *Escherichia coli* genes and evolution of stop codon usage. Journal of Theoretical Biology **168:** 461-463.

Alm, R. A., L. S. L. Ling, D. T. Moir, B. L. King, E. D. Brown *et al.*, (1999). Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. Nature, **397:** 176-180.

Altay, G., F. Bozoglu and B. Ray, (1994). Efficiency of gene-transfer by conjugation and electroporation in lactococci and pediococci. Food Microbiology**11:** 265-270.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman, (1990). Basic local alignment search tool. Journal of Molecular Biology **215:** 403-410.

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

Andersson, S. G. E., R. H. Buckingham and C. G. Kurland, (1984). Does codon composition influence ribosome function. EMBO Journal **3:** 91-94.

Andersson, S. G. E., and C. G. Kurland, (1990). Codon preferences in free living microorganisms. Microbiological Reviews **54:** 198-210.

Andersson, S. G. E., and P. M. Sharp, (1996a). Codon usage and base composition in *Rickettsia prowazekii*. Journal of Molecular Evolution, **42:** 525-536.

Andersson, S. G. E., and P. M. Sharp, (1996b). Codon usage in the *Mycobacterium tuberculosis* complex. Microbiology-UK, **142:** 915-925.

Andersson, S. G. E., A. Zomorodipour, J. O. Andersson, T. Sicheritz Ponten, U. C. M. Alsmark *et al.*, (1998). The genome sequence of *Rickettsia prowazeki* and the origin of mitochondria. Nature **396:** 133-140.

ANSI, (1978). American National Standard Fortran X3.9-1978 (FORTRAN 77). American National Standard Institute, New York.

Aota, S., T. Gojobori, F. Ishibashi, T. Maruyama and T. Ikemura, (1988). Codon usage tabulated from the GenBank genetic sequence data. Nucleic Acids Research **16:** R 315-R 402.

Aota, S., and T. Ikemura, (1986). Diversity in G+C content at the third codon position of codons in vertebrate genes and its causes. Nucleic Acids Research **14:** 6345-6355.

Aparicio, S., A. Morrison, A. Gould, J. Gilthorpe, C. Chaudhuri *et al.*, (1995). Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, *Fugu rubripes*. Proceedings of The National Academy of Sciences of The United States of America, **92:** 1684-1688.

Aquadro, C., (1992). Why is the genome variable? Insight from *Drosophila*. Trends in Genetics **8:** 355-362.

Arhin, F. F., and L. C. Vining, (1993). Cloning, nucleotide-sequence and expression in *Streptomyces lividans* and *Escherichia coli* of *pabB* from *Lactococcus lactis* subsp *lactis* NCDO-496. Journal of General Microbiology **139:** 1785-1793.

Arkov, A. L., S. V. Korolev and L. L. Kisselev, (1995). 5' contexts of *Escherichia coli* and human termination codons are similar. Nucleic Acids Research **23:** 4712-4716.

Arnold, J., A. J. Cuticchia, D. Newsome, W. Jennings and R. Ivarie, (1988). A Markov-chain analysis Mono- through hexanucleotide composition of the sense strand of yeast DNA. Nucleic Acids Research **16:** 7145-7158.

Bagnoli, F., and P. Lio, (1995). Selection, mutations and codon usage in a bacterial model. Journal of Theoretical Biology **173:** 271-281.

Bairoch, H., and B. Boeckmann, (1994). The Swiss-Prot protein sequence data bank: current status. Nucleic Acids Research **22:** 3578-3580.

Bairoch, H., and R. Apweiler, (1998). The SWISS-PROT protein sequence data bank and its suplement TrEMBL in 1998. Nucleic Acids Research **26:** 38-42.

Barril, M. J. S., S. G. Kim and C. A. Batt, (1994). Cloning and sequencing of the *Lactococcus lactis* subsp *lactis dnaK* gene using a PCR based approach. Gene **142:** 91-96.

Bartsevich, V. V., and S.V. Sestakov, (1995). The *dspA* gene product of the cyanobacterium *Synechocystis* sp. strain PCC6803 influences sensitivity to chemically different growth inhibitors and has amino acid similarity to histidine protein kinases. Microbiology **141:** 2915-2920.

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

Benson, D.A., M. Boguski, D.J. Lipman and J. Ostell, (1994). GenBank. Nucleic Acids Research **22:** 3441-3444.

Benzecri, J. P., (1992).*Correspondence analysis handbook*. Marcel Dekker, New York.

Berg, O., and M. Martelius, (1995). Synonymous substitution-rate constants in *Escherichia coli* and *Salmonella typhimurium* and their relationship to gene expression and selection pressure. Journal of Molecular Evolution **41:** 449-456.

Berger, E., (1977). Are synonymous mutations adaptively neutral? American Naturalist **111:** 606-607.

Berger, E., (1978). Pattern and chance in the use of the genetic code. Journal of Molecular Evolution **10:** 319-323.

Bernardi, G., (1989). The isochore organization of the human genome. Annual Review of Genetics **23:** 637-661.

Bernardi, G., (1993a). The isochore organization of the human genome and its evolutionary history a review. Gene **135:** 57-66.

Bernardi, G., (1993b). The vertebrate genome - isochores and evolution. Molecular Biology and Evolution **10:** 186-204.

Bernardi, G., and G. Bernardi, (1986). Compositional constraints and genome evolution. Journal of Molecular Evolution **24:** 1-11.

Bernardi, G., B. Olofsson, J. Filipski, M. Zerial, J. Salinas *et al.*, (1985). The mosaic genome of warm-blooded vertebrates. Science **228:** 953-958.

Bevan, M., I. Bancroft, H. W. Mewes, R. Martienssen and R. McCombie, (1999). Clearing a path through the jungle: progress in *Arabidopsis* genomics. Bioessays **21:** 110-120.

Bibb, M. J., P. R. Findlay and M. W. Johnson, (1984). The relationship between base composition and codon usage in bacterial genes and its use for the simple and reliable identification of protein-coding sequences. Gene **30:** 157-166.

Bird, A., (1984). CpG-rich islands and the function of DNA methylation. Nature **321:** 209-213.

Blake, J. A., J. E. Richardson, M. T. Davisson and J. T. Eppig, (1999). The mouse genome database (mgd): genetic and genomic information about the laboratory mouse. Nucleic Acids Research **27:** 95-98.

Blake, R.D., S. Hess and J. Nicholson-Tuell, (1992). The influence of nearest neighbors on the rate and pattern of spontaneous point mutations. Journal of Molecular Evolution **34:** 189-200.

Blake, R. D., and P. W. Hinds, (1984). Analysis of the codon bias in *Escherichia coli* sequences. Journal of Biomolecular Structure & Dynamics **2:** 593-606.

Blattner, F. R., G. Plunkett, C. A. Bloch, N. T. Perna, V. Burland *et al.*, (1997). The complete genome sequence of *Escherichia coli* K-12. Science **277:** 1453 (17 pages).

Bloom, B. R., (1995). A microbial minimalist. Nature **378:** 236.

Bonitz, S. G., R. Berlani, G. Coruzzi, M. Li, G. Macino *et al.*, (1980). Codon recognition rules in yeast mitochondria. Proceedings of The National Academy of Sciences of The United States of America **77:** 3167-3170.

Borodovsky, M., E. V. Koonin and K. E. Rudd, (1994a). New Genes in Old Sequence - a Strategy for Finding Genes in the Bacterial Genome. **19:** 309-313.

Borodovsky, M., and J. D. McIninch, (1993). Genmark - Parallel Gene Recognition For Both DNA Strands. Computers and Chemistry **17:** 123-133.

Borodovsky, M., J. D. McIninch, E. V. Kounin, K. E. Rudd, C. Medigue *et al.*, (1995). Detection of new genes in a bacterial genome using Markov models for 3 gene classes. Nucleic Acids Research **23:** 3554-3562.

Borodovsky, M., K. E. Rudd and E. V. Kounin, (1994b). Intrinsic and Extrinsic Approaches for Detecting Genes in a Bacterial Genome. Nucleic Acids Research **22:** 4756-4767.

Bowman, S., C. Churcher, K. Badcock, D. Brown, T. Chillingworth *et al.*, (1997). The nucleotide sequence of *Saccharomyces cerevisiae* chromosome XIII. Nature, **387:** 90-93.

Bradnam, K.R., C. Seoighe, P. M. Sharp and K. H. Wolfe, (1999). G+C content variation along and among *Saccharomyces cerevisiae* chromosomes. Molecular Biology Evolution **16:** 666-675.

Breidt, F., and G. C. Stewart, (1987). Nucleotide and deduced amino acid sequence of the *Staphylococcus aureus* phospho-β-galactosidase gene. Applied Environmental Microbiology **53:** 969-973.

Brown, A., (1989). Messenger RNA stability on yeast. Yeast **5:** 239-257.

Brown, C. M., M. E. Dalphin, P. A. Stockwell and W. P. Tate, (1993). The translational termination signal database. Nucleic Acids Research **21:** 3119-3123.

Brown, C. M., P. A. Stockwell, M. E. Dalphin and W. P. Tate, (1994). The translational termination signal database (Transterm) now also includes initiation contexts. Nucleic Acids Research **22:** 3620-3624.

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

Bulmer, M., (1988). Are codon usage patterns in unicellular organisms determined by selection-mutation balance. Journal of Evolutionary Biology **1:** 15-26.

Bulmer, M., (1990). The effect of context on synonymous codon usage in genes with low codon usage bias. Nucleic Acids Research **18:** 2869-2873.

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

Bult, C. J., O. White, G. J. Olsen, L. X. Zhou, R. D. Fleischmann *et al.*, (1996). Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. Science **273:** 1058-1073.

Buvinger, W., K. Lampel, R. Bojanowski and M. Riley, (1984). Location and analysis of nucleotide sequences at one end of a putative *lac* transposon in *Escherichia coli*. Journal of Bacteriology **159:** 618-623.

Caccio, S., P. Perani, S. Saccone, F. Kadi and G. Bernardi, (1994). Single copy sequence homology among the GC richest isochores of the genomes from warm blooded vertebrates. Journal of Molecular Evolution **39:** 331-339.

Cancilla, M. R., B. E. Davidson, A. J. Hillier, N. Y. Nguyen and J. Thompson, (1995a). The *Lactococcus lactis* triosephosphate isomerase gene, *tpi*, is monocistronic. Microbiology-UK **141:** 229-238.

Cancilla, M. R., A. J. Hillier and B. E. Davidson, (1995b). *Lactococcus lactis* glyceraldehyde-3-phosphate dehydrogenase gene, *gap* - further evidence for strongly biased codon usage in glycolytic pathway genes. Microbiology-UK, **141:** 1027-1036.

Chauvat, C., M. Poncelet and F. Chauvat, (1997). Three insertion sequences from the cyanobacterium *Synechocystis* PCC6803 support the occurrence of horizontal DNA transfer among bacteria. Gene **195:** 257-266.

Chassy, B. M., and C. A. Alpert, (1989). Molecular organisation of the plasmid encoded lactose-PTS of *Lactobacillus casei*. FEMS Microbiology Reviews **63:** 157-166.

Chavancy, G., A. Chevallier, A. Fournier and J.-P. Garel, (1979). Adaptation of iso-tRNA concentration to mRNA codon frequency in the eukaryote cell. Biochimie **61:** 71-78.

Chavancy, G., and J.-P. Garel, (1981). Does quantitative tRNA adaptation to codon content in mRNA optimise the ribosomal translation efficiency? Proposal for a translational system model. Biochimie **63:** 187-195.

Chen, G. F. T., and M. Inouye, (1990). Suppression of the negative effect of minor Arginine codons on gene-expression - preferential usage of minor codons within the 1st 25 codons of the *Escherichia coli* genes. Nucleic Acids Research **18:** 1465-1473.

Chen, G. F. T., and M. Inouye, (1994). Role of the AGA/AGG codons, the rarest codons in global gene expression in *Escherichia coli*. Genes & Development **8:** 2641-2652.

Chopin, A., (1993). Organization and regulation of genes for amino-acid biosynthesis in lactic-acid bacteria. FEMS Microbiology Reviews **12:** 21-38.

Clarke, B. C., (1970). Darwinian evolution of proteins. Science **168:** 1009-1011.

Cole, S. T., R. Brosch, J. Parkhill, T. Garnier, C. Churcher *et al.*, (1998). Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence (vol 393, pg 537, 1998). Nature **396:** 190-198.

Collins, M. D., C. Ash, J. A. E. Farrow, S. Wallbanks and A. M. Williams, (1989). 16S ribosomal ribonucleic acis sequence analysis of lactococci and related taxa. Description of Vagococcus fluvialis gen, nov. sp. nov,. Journal of Applied Bacteriology **67:** 453-460.

Collins, R. F., M. Roberts and D. A. Phoenix, (1995). Codon bias in *Escherichia coli* may modulate translation initiation. Biochemical Society Transactions **23:** S 76.

Craigen, W. J., R. G. Cook, W. P. Tate and C. T. Caskey, (1985). Bacterial peptide chain release factors: Conserved primary structure and possible frameshift regulation of release factor 2. Proceedings of The National Academy of Sciences of The United States of America **82:** 3616-3620.

Crick, F. H. C., (1968). The origin of the Genetic Code. Journal of Molecular Biology **38:** 367-379.

Croux, C., and J. L. Garcia, (1992). Reconstruction and expression of the autolytic gene from *Costridium aetobutylicum* ATCC-824 in *Escherichia-coli*. FEMS Microbiology Letters **95:** 13-20.

Curran, J. F., (1995). Decoding with the A-I wobble pair is inefficient. Nucleic Acids Research **23:** 683-688.

Curran, J. F., and M. Yarus, (1988). Rates of AA-tRNA selection at 29 sense codons *in vivo*. Journal of Molecular Biology **209:** 65-77.

de Smit, M.H., J. van Duin, P. H. van Knippenberg and H.G. van Eijk, (1994). CCC.UGA: a new site of ribosomal frameshifting in *Escherichia coli*. Gene **143:** 43-47.

de Smit, M. H., and J. van Duin, (1990a). Control of prokaryotic translational initiation by mRNA secondary structure, vol. 38, pp. 1-35 in *Progress in Nucleic Acid Research and Molecular Biology*. Academic Press Inc.

de Smit, M. H., and J. van Duin, (1990b). Secondary structure of the ribosome binding-site determines translational efficiency - a quantitative-analysis. Proceedings of The National Academy of Sciences of The United States of America **87:** 7668-7672.

de Smit, M. H., and J. van Duin, (1994). Control of translation by messenger RNA secondary structure in *Escherichia coli* a quantitative analysis of literature data. Journal of Molecular Biology **244:** 144-150.

de Vos, W. M., I. Boerrigter, R. J. van Rooyen, B. Reiche and W. Hengstenberg, (1990). Characterization of the lactose-specific enzymes of the phosphotransferase system in *Lactococcus lactis*. Journal of Biological Chemistry **265:** 22554-22560.

de Vos, W. M., and E. E. Vaughan, (1994). Genetics of lactose utilization in lactic acid bacteria. FEMS Microbiology Reviews **15:** 217-237.

Deckert, G., P. V. Warren, T. Gaasterland, W. G. Young, A. L. Lenox *et al.*, (1998). The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. Nature **392:** 353-358.

Delorme, C., J. J. Godon, S. D. Ehrlich and P. Renault, (1994). Mosaic structure of large regions of the *Lactococcus lactis* subsp *cremoris* chromosome. Microbiology-UK, **140:** 3053-3060.

Deschavanne, P., and J. Filipski, (1995). Correlation of GC content with replication timing and repair mechanisms in weakly expressed *Escherichia coli* genes. Nucleic Acids Research **23:** 1350-1353.

Dong, H. J., L. Nilsson and C. G. Kurland, (1995). Gratuitous overexpression of genes in *Escherichia coli* leads to growth-inhibition and ribosome destruction. Journal of Bacteriology **177:** 1497-1504.

D'onofrio, G., and G. Bernardi, (1992). A universal compositional correlation among codon positions. Gene **110:** 81-88.

Dufour, A., D. Thuault, A. Boulliou, C. M. Bourgeois and J. P. Lepennec, (1991). Plasmid-encoded determinants for bacteriocin production and immunity in a *Lactococcus lactis* strain and purification of the inhibitory peptide. Journal of General Microbiology **137:** 2423-2429.

Dujon, B., K. Albermann, M. Aldea, D. Alexandraki, W. Ansorge *et al.*, (1997). The nucleotide sequence of *Saccharomyces cerevisiae* chromosome XV. Nature, **387:** 98-102.

Dujon, B., D. Alexandraki, B. Andre, W. Ansorge, V. Baladron *et al.*, (1994). Complete DNA sequence of yeast Chromosome XI. Nature, **369:** 371-378.

Eaton, T., C. Shearman and M. Gasson, (1993). Cloning and sequence analysis of the *dnaK* gene region of *Lactococcus lactis* subsp *lactis*. Journal of General Microbiology **139:** 3253-3264.

Efstathiou, J. D., and L. L. McKay, (1976). Plasmids in *Streptococcus lactis*: evidence that lactose metabolism and proteinsase activity are plasmid linked. Applied Environmental Microbiology **32:** 38-44.

Ehrenberg, M., and C. G. Kurland, (1984). Costs of accuracy determined by a maximal growth-rate constraint. Quarterly Reviews of Biophysics **17:** 45-82.

Eikmanns, B. J., (1992). Identification, sequence analysis, and expression of *a Corynebacterium glutamicum* gene cluster encoding the three glycolytic enzymes glyceraldehyde-3-phosphate dehydrogenase, 3-phosphoglycerate kinase, and triose phosphate isomerase. Journal of Bacteriology **174:** 6076-6086.

Ellis, J., H. Griffin, D. Morrison and A. M. Johnson, (1993). Analysis of dinucleotide frequency and codon usage in the phylum Apicomplexa. Gene **126:** 163-170.

Ellis, J., D. A. Morrison and B. Kalinna, (1995). Comparison of the patterns of codon usage and bias between *Brugia*, *Echinococcus, Onchocerca* and *Schistosoma* species. Parasitology Research **81:** 388-393.

Ellis, J. T., and D. A. Morrison, (1995). *Schistosoma mansoni* patterns of codon usage and bias. Parasitology **110:** 53-60.

Ellis, J. T., D. A. Morrison, D. Avery and A. M. Johnson, (1994). Codon usage and bias among individual genes of the Coccidia and Piroplasms. Parasitology **109:** 265-272.

Elton, B., G. J. Russell and J. Subak-Sharpe, (1976). Doublet frequencies and codon weighting in the DNA of *Escherichia coli.* Journal of Molecular Evolution **8:** 117-135.

Emilsson, V., and C. G. Kurland, (1990a). Growth rate dependence of transfer-RNA abundance in *Escherichia coli*. EMBO Journal **9:** 4359-4366.

Emilsson, V., and C. G. Kurland, (1990b). Growth-rate dependence of global amino acid composition. Biochimica et Biophysica Acta **1050:** 248-251.

Emilsson, V., A. K. Naslund and C. G. Kurland, (1993). Growth-rate dependent accumulation of 12 transfer-RNA species in *Escherichia coli*. Journal of Molecular Biology **230:** 483-491.

Emmert, D. B., P. J. Stoehr, G. Stoesser and G. N. Cameron, (1994). The European Bioinformatics Institute (EBI) databases. Nucleic Acids Research **22:** 3445-3449.

Etzold, T., and P. Argos, (1993). SRS an indexing and retrieval tool for flat data libraries. Computer Applications for the Biosciences **9:** 49-57.

Eyre-Walker, A., (1991). An analysis of codon usage in mammals - selection or mutation bias. Journal of Molecular Evolution **33:** 442-449.

Eyre-Walker, A., (1994a). DNA mismatch repair and synonymous codon evolution in mammals. Molecular Biology and Evolution **11:** 88-98.

Eyre-Walker, A., (1994c). Synonymous substitutions are clustered in Enterobacterial genes. Journal of Molecular Evolution **39:** 448-451.

Eyre-Walker, A., (1995a). The distance between *Escherichia coli* genes is related to gene expression levels. Journal of Bacteriology, **177:** 5368-5369.

Eyre-Walker, A., (1995b). Does very short patch (VSP) repair efficiency vary in relation to gene-expression levels. Journal of Molecular Evolution **40:** 705-706.

Eyre-Walker, A., and M. Bulmer, (1993). Reduced synonymous substitution rate at the start of Enterobacterial genes. Nucleic Acids Research **21:** 4599-4603.

Eyre-Walker, A., and M. Bulmer, (1995). Synonymous substitution rates in Enterobacteria. Genetics **140:** 1407-1412.

Feldmann, H., M. Aigle, G. Aljinovic, B. Andre, M. C. Baclet *et al.*, (1994). Complete DNA sequence of yeast Chromosome II. EMBO Journal **13:** 5795-5809.

Fennoy, S. L., and J. Baileserres, (1993). Synonymous codon usage in *Zea mays* nuclear genes is varied by levels of C-ending and G-ending codons. Nucleic Acids Research **21:** 5294-5300.

Fichant, G., and C. Gautier, (1987). Statistical methods for prediction of protein coding regions in nucleic acids sequences. Computer Applications for the Biosciences **3:** 287-295.

Fiers, W., R. Contreras, F. Duerinck, G. Haegeman, J. Iserentant *et al.*, (1976). Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. Nature **260:** 500-507.

Fiers, W., R. Contreras, F. Duerinck, G. Haegmean, J. Merregaert *et al.*, (1975). A-protein of bacteriophage MS2. Nature **256:** 273-278.

Filipski, J., (1987). Correlation between molecular clock ticking, codon usage, fidelity of DNA repair, chromosome-banding and chromatin compactness in germline cells. FEBS Letters **217:** 184-186.

Fitch, D. H. A., and L. D. Strausbaugh, (1993). Low codon bias and high-rates of synonymous substitution in *Drosophila hydei* and *Drosophila melanogaster* histone genes. Molecular Biology and Evolution **10:** 397-413.

Fitch, W., (1976). Is there selection against wobble in codon-anticodon pairing ? Science **194:** 1173-1174.

Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness *et al.*, (1995). Whole genome random sequencing and assembly of *Haemophilus influenzae* RD. Science **269:** 496-512.

Forsburg, S. L., (1994). Codon usage table for *Schizosaccharomyces pombe*. Yeast **10:** 1045-1047.

Forsdyke, D. R., (1995a). Relative roles of primary sequence and (G+C)% in determining the hierarchy of frequencies of complementary trinucleotide pairs in DNAs of different species. Journal of Molecular Evolution **41:** 573-581.

Forsdyke, D. R., (1995b). Sense in Antisense? Journal of Molecular Evolution **41:** 582-586.

Fraser, C. M., J. D. Gocayne, O. White, M. D. Adams, R. A. Clayton *et al.*, (1995). The minimal gene complement of *Mycoplasma genitalium*. Science, **270:** 397-403.

Fraser, C. M., S. J. Norris, C. M. Weinstock, O. White, G. G. Sutton *et al.*, (1998). Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. Science **281:** 375-388.

Fraser, C. M., S. Casjens, W.M. Huang, G.G. Sutton, R. Clayton *et al.*, (1997). Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. Nature **390:** 580-586.

Freire Picos, M. A., M. I. Gonzalez Siso, E. Rodriguez Belmonte, A. M. Rodriguez Torres, E. Ramil *et al.*, (1994). Codon usage in *Kluyveromyces lactis* and in yeast cytochrome c-encoding genes. Gene **139:** 43-49.

Friedberg, E. C., A. J. Bardwell, L. Bardwell, Z. Wang and G. Dianov, (1994). Transcription and nucleotide excision repair reflections, considerations and recent biochemical insights. Mutation Research **307:** 5-14.

Frohlich, D. R., and M. A. Wells, (1994). Codon usage patterns among genes for *Lepidopteran hemolymph* proteins. Journal of Molecular Evolution **38:** 476-481.

Fsihi, H., and S. T. Cole, (1995). The *Mycobacterium leprae* genome systematic sequence analysis identifies key catabolic enzymes, ATP dependent transport systems and a novel polA locus associated with genomic variability. Molecular Microbiology, **16:** 909-919.

Galas, D., and T. Smith, (1984). The relationship between codon boundaries and multiple reading-frame preferences: coding organization of bacterial insertion sequences. Molecular Biology and Evolution **11:** 260-268.

Ganoza, M. C., and B. G. Louis, (1994). Potential secondary structure at the translational start domain of eukaryotic and prokaryotic messenger-RNAs. Biochimie **76:** 428-439.

Gardner, M. J., H. Tettelin, D. J. Carucci, L. M. Cummings, L. Aravind *et al.*, (1998). Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. Science, **282:** 1126-1132.

Garel, J.-P., (1974). Functional adaptation of tRNA population. Journal of Theoretical Biology **43:** 211-225.

Garvie, E. I., and J. A. E. Farrow, (1982). *Streptococcus lactis* subsp. *cremoris* (Orla-Jensen) comb. nov. and *Streptococcus lactis* subsp. *diacetilactis* (Matuszewski et al.) nov. rev., comb. nov. International Journal of Systematic Bacteriology **32:** 453-455.

GCG (1994). Programme Manual for the Wisconsin Package, Version 8, University of Wisconsin.

George, D.G., W.C. Barker, H.-W. Mewes, F. Pfeiffer and A. Tsugita, (1994). The PIR international protein sequence database. Nucleic Acids Research **22:** 3569-3573.

Gesteland, R. F., R. B. Weiss and J. F. Atkins, (1992). Recoding - reprogrammed genetic decoding. Science **257:** 1640-1641.

Gething, M. J., and J. Sambrook, (1992). Protein folding in the cell. Nature **355:** 33-45.

Gharbia, S. E., J. C. Williams, D. M. A. Andrews and H. N. Shah, (1995). Genomic clusters and codon usage in relation to gene-expression in oral gram-negative anaerobes. Anaerobe, **1:** 239-262.

Gireesh, T., B. E. Davidson and A. J. Hillier, (1992). Conjugal transfer in *Lactococcus lactis* of a 68-kilobase-pair chromosomal fragment containing the structural gene for the peptide bacteriocin nisin. Applied and Environmental Microbiology **58:** 1670-1676.

Godon, J. J., C. Delmore, S. D. Ehrlich and P. Renault, (1992). Divergence of genomic sequences between *Lactococcus lactis* subsp. lactis and *Lactococcus lactis* subsp. cremoris. Applied and Environmental Microbiology **58:** 4045-4047.

Goffeau, A., R. Aert, M. L. Agostini-Carbone et. al., (1997). The Yeast Genome Directory. Nature **387:** 5-105.

Goldman, E., A. H. Rosenberg, G. Zubay and F. W. Studier, (1995). Consecutive low usage leucine codons block translation only when near the 5' end of a message in *Escherichia coli*. Journal of Molecular Biology **245:** 467-473.

Goldman, N., and Z. H. Yang, (1994). Codon based model of nucleotide substitution for protein coding DNA sequences. Molecular Biology and Evolution **11:** 725-736.

Gouy, M., (1987). Codon contexts in Enterobacterial and Coliphage genes. Molecular Biology and Evolution **4:** 426-444.

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

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

Grantham, R., C. Gautier and M. Gouy, (1980a). Codon frequencies in 119 genes confirm consistent choices of degenerate base according to genome type. Nucleic Acids Research **8:** 1892-1912.

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

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

Grantham, R., T. Greenland, S. Louail, D. Mouchiroud, J. Prato *et al.*, (1985). Molecular evolution of viruses as seen by nucleic acids sequence study. Bulletin Institute Pasteur **83:** 95-148.

Greenacre, M. J., (1984).*Theory and applications of correspondence analysis.* Academic Press, London.

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

Groisman, E. A., M. H. Saier and H. Ouchmann, (1992). Horizontal transfer of a phosphatase gene as evidence for mosaic structure of the *Salmonella* genome. EMBO Journal **11:** 1309-1316.

Groisman, E. A., M. A. Sturmoski, F. R. Solomon, R. Lin and H. Ochmann, (1993). Molecular, functional and evolutionary analysis of sequence specific to *Salmonella*. Proceedings of The National Academy of Sciences of The United States of America **90:** 1033-1037.

Grosjean, H., and W. Fiers, (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.

Gu, X., and W. H. Li, (1994). A model for the correlation of mutation rate with GC content and the origin of GC rich isochores. Journal of Molecular Evolution **38:** 468-475.

Gursky, Y.G., and R.S. Beabealashvilli, (1994). The increase in gene expression induced by introduction of rare codons into the C terminus of the template. Gene **148:** 15-21.

Gutierrez, G., J. Casadesus, J. L. Oliver and A. Marin, (1994). Compositional heterogeneity of the *Escherichia coli* genome - a role for VSP repair. Journal of Molecular Evolution **39:** 340-346.

Gutman, G. A., and G. W. Hatfield, (1989). Nonrandom utilization of codon pairs in *Escherichia coli*. Proceedings of The National Academy of Sciences of The United States of America **86:** 3699-3703.

Hamada, S., and H. D. Slade, (1980). Biology, immunology, and cariogenicity of *Streptococcus mutans*. Microbiology Reviews **44:** 331-384.

Hanawalt, P. C., (1991). Heterogeneity of DNA repair at the gene level. Mutational Research **247:** 203-211.

Harris, L. J., H. P. Fleming and T. R. Klaenhammer, (1992). Developments in nisin research. Food Research International **25:** 57-66.

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

Hasegawa, M., T. Yasunaga and T. Miyata, (1979). Secondary structure of MS2 phage RNA and bias in code word usage. Nucleic Acid Usage **7:** 2073-2079.

Hastings, K., and C. Emerson, (1983). Codon usage on muscle genes and liver genes. Journal of Molecular Evolution **19:** 214-218.

Hengstenberg, W., J. B. Egan and M. L. Morse, (1967). Carbohydrate transport in *Staphylococcus aureus*. V. The accumulation of phosphorylated carbohydrate derivates, and evidence for a new enzyme splitting lactose phosphate. Proceedings of the National Academy of Sciences of The United States of America. **58:** 274-279.

Herrick, D., R. Parker and A. Jacobson, (1980). Identification and comparison of stable and unstable mRNAs in *Saccharomyces cerevisiae*. Molecular Cell Biology **10:** 2269-2284.

Higgins, D., (1992). Sequence ordinations: a multivariate analysis approach to analysing large sequence data sets. Computer Applications for the Biosciences **8:** 15-22.

Hill, M. O., (1974). Correspondence analysis: a neglected multivariate method. Applied Statistics **23:** 340-354.

Himmelreich, R., H. Hilbert, H. Plagens, E. Pirkl, B. Li *et al.*, (1996). Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. Nucleic Acids Research **24:** 4420-4449.

Hirel, P. H., J. M. Schmitter, P. Dessen, G. Fayat and S. Blanquet, (1989). Extent of N-terminal Methionine excision from *Escherichia coli* proteins is governed by the side chain length of the penultimate amino acid. Proceedings of The National Academy of Sciences of The United States of America **86:** 8247-8251.

Hirosawa, M., K. Isono, W. S. Hayes and M. Borodovsky, (1997). Gene identification and classification in the *Synechocystis* genomic sequence by recursive gene mark analysis. DNA Sequence **8:** 17-29.

Hoekema, A., R. A. Kastelein, M. Vasser and H. A. Deboer, (1987). Codon replacement in the *pgk1* gene of *Saccharomyces cerevisiae* - experimental approach to study the role of biased codon usage in gene-expression. Molecular and Cellular Biology **7:** 2914-2924.

Holm, L., (1986). Codon usage and gene expression. Nucleic Acids Research **14:** 3075-3087.

Honeyman, A. L., and R. Curtiss, (1993). Isolation, characterization and nucleotide sequence of the *Streptococcus mutans* lactose-specific enzyme II *(lacE*) gene of the PTS and the phospho-beta-galactosidase (*lacG*) gene. Journal of General Microbiology **139:** 2685-2694.

Hua, Z., H. Wang, D. Chen, Y. Chen and D. Zhu, (1994). Enhancement of expression of human granulocyte macrophage stimulating factor by *argU* gene product in *Escherichia coli*. Biochemistry Molecular Biology **32:** 537-543.

Ikemura, T., (1981a). 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* system. Journal of Molecular Biology **151:** 389-409.

Ikemura, T., (1981b). Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons of its protein genes. Journal of Molecular Biology **146:** 1-21.

Ikemura, T., (1982). Correlation between the abundance of yeast tRNAs 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. Journal of Molecular Biology **158:** 573-597.

Ikemura, T., (1985). Codon usage and transfer-RNA content in unicellular and multicellular organisms. Molecular Biology and Evolution **2:** 13-34.

Ikemura, T., and H. Ozeki, (1982). Codon usage and transfer RNA contents: organism specific codon choice patterns in reference to the isoacceptor contents. Cold Spring Harbor Symposium Quantitative Biology **47:** 1087-1097.

Ikemura, T., and K. Wada, (1991). Evident diversity of codon usage patterns of human genes with respect to chromosome banding patterns and chromosome numbers relation between nucleotide-sequence data and cytogenetic data. Nucleic Acids Research **19:** 4333-4339.

Isacchi, A., G. Bernardi and G. Bernardi, (1993). Compositional compartmentalization of the nuclear genomes of *Trypanosoma brucei* and *Trypanosoma equiperdum*. FEBS Letters **335:** 181-183.

Ivanov, I., R. Alexandrova, B. Dragulev, A. Saraffova and M. G. Abouhaidar, (1992). Effect of tandemly repeated AGG triplets on the translation of *cat* messenger RNA in *Escherichia coli*. FEBS Letters **307:** 173-176.

Jacq, C., J. Altmorbe, B. Andre, W. Arnold, A. Bahr *et al.*, (1997). The nucleotide sequence of *Saccharomyces cerevisiae* chromosome IV. Nature, **387:** 75-78.

Jacques, N., and M. Dreyfus, (1990). Translation initiation in *Escherichia coli* - old and new questions. Molecular Microbiology **4:** 1063-1067.

Jarvis, A. W., and B. D. W. Jarvis, (1981). Deoxyribonucleic acid homology among lactic streptococci. Applied Environmental Microbiology **41:** 77-83.

Jinks-Robertson, S., and N. Nomura, (1987). *Ribosomes and tRNA*. American Society for Microbiology, Washington DC.

Johnston, M., S. Andrews, R. Brinkman, J. Cooper, H. Ding *et al.*, (1994). Complete nucleotide sequence of *Saccharomyces cerevisiae* Chromosome VIII. Science **265:** 2077-2082.

Kadi, F., D. Mouchiroud, G. Sabeur and G. Bernardi, (1993). The compositional patterns of the avian genomes and their evolutionary implications. Journal of Molecular Evolution **37:** 544-551.

Kafatos, F., A. Efstratiadis and B. Forget, (1977). Molecular evolution of human and rabbit beta-globin mRNAs. Proceedings of The National Academy of Sciences of The United States of America **74:** 5618-5622.

Kagawa, Y., H. Nojima, N. Nukiwa, M. Ishizuka and T. Nakajima, (1984). High guanine plus cytosine content in the third codon letter of an extreme thermophile. Journal of Biological Chemistry **259:** 2956-2960.

Kalinna, B. H., and D. P. McManus, (1994). Codon usage in *Echinococcus*. Experimental Parasitology **79:** 72-76.

Kane, J. F., (1995). Effects of rare codon clusters on high-level expression of heterologous proteins in *Escherichia coli*. Current Opinion In Biotechnology **6:** 494-500.

Kane, J. F., B. Violand, D. Curran, N. Staten, K. Duffin *et al.*, (1993). Novel in frame two codon translational hop during synthesis of bovine placental lactogen in a recombinant strain of *E. coli*. Nucleic Acids Research **20:** 6707-6712.

Kaneko, T., T. Matsubayashi, M. Sugita and M. Sugiura, (1996a). Physical and gene maps of the unicellular cyanobacterium *Synechococcus* sp strain PCC6301 genome. Plant Molecular Biology, **31:** 193-201.

Kaneko, T., S. Sato, H. Kotani, A. Tanaka, E. Asamizu *et al.*, (1996b). Sequence analysis of the genome of the unicellular *Cyanobacterium Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and the assignment of potential protein-coding regions. DNA Research **3:** 109-136.

Karlin, S., and L. R. Cardon, (1994). Computational DNA sequence analysis. Annual Review of Microbiology **48:** 619-654.

Karlin, S., J. Mrazek and A. M. Campbell, (1997). Compositional biases of bacterial genomes and evolutionary implications. Journal of Bacteriology **179:** 3899-3913.

Kaufman, L., and P. J. Rousseeuw, (1990). *Finding groups in data: an introduction to data analysis*. Wiley Interscience, New York.

Kernighan, B. W., and D. M. Ritchie, (1988). *The C programming language*. Prentice-Hall, Englewood Cliffs, NJ.

Kerr, A. R. W., J. F. Peden and P.M. Sharp, (1997). Systematic base composition variation around the genome of *Mycoplasma genitalium*, but not *Mycoplasma pneumoniae*. Molecular Microbiology **25:** 1177-1179.

Kimura, M., (1968). Evolutionary rate at the molecular level. Nature **217:** 624-626.

Kimura, M., (1977). Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature **267:** 275-276.

Kimura, M., (1981). Possibility of extensive neutral evolution under stabilising selection with special reference to non-random usage of synonymous codons. Proceedings of the National Academy of Sciences of The United States of America **78:** 5773-5777.

Kimura, M., (1983).*The Neutral Theory of Molecular evolution*. Cambridge University Press, Cambridge.

King, J. L., and T. H. Jukes, (1969). Non-Darwinian evolution. Science **164:** 788-798.

Klaenhammer, T. R., (1993). Genetics of bacteriocins produced by lactic-acid bacteria. FEMS Microbiology Reviews **12:** 39-86.

Klein, J. R., B. Henrich and R. Plapp, (1994). Cloning and nucleotide sequence analysis of the *Lactobacillus delbrueckii* ssp *lactis* DSM7290 cysteine aminopeptidase gene *pepC*. FEMS Microbiology Letters **124:** 291-299.

Klenk, H. P., R. A. Clayton, J. F. Tomb, O. White, K. E. Nelson *et al.*, (1997). The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. Nature **390:** 364-379

Kliman, R. M., and J. Hey, (1993). Reduced natural-selection associated with low recombination in *Drosophila melanogaster*. Molecular Biology and Evolution **10:** 1239-1258.

Kliman, R. M., and J. Hey, (1994). The effects of mutation and natural-selection on codon bias in the genes of *Drosophila*. Genetics **137:** 1049-1056.

Kloos, W. E., and K. H. Schleifer, (1986). Genus IV *Staphylococcus*, pp. 1013-1019 in *Bergey's Manual of Systematic Bacteriology*, edited by P. H. Sneath, N. S. Mair, M. E. Sharpe and J. G. Holt. Williams and Wilkins, London.

Koivula, T., and H. Hemila, (1991). Nucleotide sequence of a *Lactococcus lactis* gene-cluster encoding adenylate kinase, initiation factor 1 and ribosomal proteins. Journal of General Microbiology **137:** 2595-2600.

Kok, J., K. J. Leenhouts, A. J. Haandrikman and A. M. Ledeboer, (1988). Nucleotide sequence of the cell wall proteinase gene of *Streptococcus cremoris* WG2. Applied Environmental Microbiology **54:** 231-238.

Kolter, R., and C. Yanofsky, (1982). Attenuation in amino acid synthetic operons. Annual Review of Genetics **16:** 113-134.

Komine, Y., T. Adaki, H. Inokuchi and H. Ozeki, (1990). Genomic organization and physical mapping of the transfer-RNA Genes in *Escherichia coli* K12. Journal of Molecular Biology **212:** 579-598.

Konigsberg, W., and G. Godson, (1983). Evidence for use of rare codons in the *dnaG* gene and other regulatory genes of *Escherichia coli.* Proceedings of The National Academy of Sciences of The United States of America **80:** 687-691.

Kotani, H., and S. Tabata, (1998). Lessons from sequencing of the genome of a unicellular cyanobacterium, *Synechocystis* sp. PCC6803. Annual Review of Plant Physiology and Plant Molecular Biology, **49:** 151-171.

Krishnaswamy, S., and S. Shanmugasundaram, (1995). Codon analysis of cyanobacterial genes. Current Science **69:** 182-185.

Krogh, A., I. S. Mian and D. Haussler, (1994). A hidden Markov model that finds genes in *Escherichia coli* DNA. Nucleic Acids Research, **22:** 4768-4778.

Kumar, P. A., and R. P. Sharma, (1995). Codon usage in brassica genes. Journal of Plant Biochemistry and Biotechnology **4:** 113-115.

Kunst, F., N. Ogasawara, I. Moszer, A. M. Albertini, G. Alloni *et al.*, (1997). The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. Nature **390:** 249-256.

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

Kurland, C. G., (1993). Major codon preference theme and variations. Biochemical Society Transactions **21:** 841- 846.

Kushiro, A., M. Shinizu and K. Tomita, (1987). Molecular cloning and sequence determination of the *tuf* gene coding for the elongation factor Tu of *Thermus thermophilus*. European Journal of Biochemistry **170:** 93-98.

Kyte, J., and R. Doolittle, (1982). A simple method for displaying the hydropathic character of a protein. Journal of Molecular Biology **157:** 105-132.

Lafay, B., A. T. Lloyd, M. J. McLean, K. M. Devine, P. M. Sharp *et al.*, (1999). Proteome composition and codon usage in spirochaetes: species -specific and DNA strand-specific mutational biases. Nucleic Acids Research **27:** 1642-1649.

Langer, T., C. Lu, H. Echols, J. Flanaghan, M. K. Hayer *et al.*, (1992). Successive action of DnaK, DnaJ and GroEL along the pathway of chaperone mediated protein folding. Nature **356:** 683-689.

Larsen, B., J. F. Peden, S. Matsufuji, T. Matsufuji, K. Brady *et al.*, (1996). Upstream stimulators for recoding. Biochemistry Cell Biology **73**: 1123-1129.

Lebart, L., A. Morineau and K. Warwick, (1984). *Multivariate descriptive statistical analysis: Correspondence analysis and related techniques for large matrices*. John Wiley & Sons, New York.

Li, W., (1993). Unbiased estimations of the rates of synonymous and nonsynonymous substitution. Journal of Molecular Evolution **36:** 96-99.

Li, W. -H., (1987). Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. Journal of Molecular Evolution **24:** 337-345.

Li, W. -H., C. I. Wu and C. C. Luo, (1985a). A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Molecular Biology and Evolution **2:** 150-174.

Li, W.-.-H, and D. Graur, (1991). *Fundamentals of Molecular Evolution*. Sinauer Associates, Sunderland MA.

Li, W.--H., C.-C. Luo and C.-I. Wu, (1985b). *Evolution of DNA sequences.* Plenum.

Lio, P., S. Ruffo and M. Buiatti, (1994). 3rd codon G+C periodicity as a possible signal for an internal selective constraint. Journal of Theoretical Biology **171:** 215-223.

Llanos, R. M., C. J. Harris, A. J. Hillier and B. E. Davidson, (1993). Identification of a novel operon in *Lactococcus lactis* encoding 3 enzymes for lactic acid synthesis phosphofructokinase, pyruvate-kinase, and lactate-dehydrogenase. Journal of Bacteriology **175:** 2541-2551.

Llanos, R. M., A. J. Hillier and B. E. Davidson, (1992). Cloning, nucleotide sequence, expression, and chromosomal location of *ldh*, the gene encoding l-(+)-lactate dehydrogenase, from *Lactococcus lactis*. Journal of Bacteriology **174:** 6956-6964.

Lloyd, A. T., and P. M. Sharp, (1991). Codon usage in *Aspergillus nidulans*. Molecular & General Genetics **230:** 288-294.

Lloyd, A. T., and P. M. Sharp, (1992a). CODONS - a microcomputer program for codon usage analysis. Journal of Heredity **83:** 239-240.

Lloyd, A. T., and P. M. Sharp, (1992b). Evolution of codon usage patterns - the extent and nature of divergence between *Candida albicans and Saccharomyces cerevisiae*. Nucleic Acids Research **20:** 5289-5295.

Lloyd, A. T., and P. M. Sharp, (1993). Synonymous codon usage in *Kluyveromyces lactis*. Yeast **9:** 1219-1228.

Lobry, J. R., and C. Gautier, (1994). Hydrophobicity, expressivity and aromaticity are the major trends of amino acid usage in 999 *Escherichia coli* chromosome encoded genes. Nucleic Acids Research **22:** 3174-3180.

Long, M. Y., and J. H. Gillespie, (1991). Codon usage divergence of homologous vertebrate genes and codon usage clock. Journal of Molecular Evolution **32:** 6-15.

Loshon, C. A., F. Tovar-Rojo, S. E. Goldrick and P. Setlow, (1989). The expression of a highly expressed *Bacillus subtilis* gene is not reduced by introduction of multiple codons normally not present. FEMS Microbiology Letters **65:** 59-64.

Maenpaa, P., and M. Bernfield, (1975). Subcellular distribution of seryl-transfer RNA during estrogen induced phosvitin synthesis and specificity of the estrogen. Biochemistry **14:** 4820-4826.

Malakhov, M. P., and V. E. Semenenko, (1994). Codon usage in genes of Cyanobacterium *Synechocystis* PCC6803. Russian Journal of Plant Physiology, **41:** 141-146.

Malumbres, M., J. A. Gil and J. F. Martin, (1993). Codon preference in Corynebacteria. Gene **134:** 15-24.

Mantegna, R., S. Buldyrev, A. Goldberger, S. Havlin, C.-K. Peng *et al.*, (1994). Linguistic features of noncoding DNA. Physical Review Letters **73:** 3169-3172.

Marshall, E., (1999). Human genome project - Sequencers endorse plan for a draft in 1 year. Science **284:** 1439.

Martin, C., C. Sibold and R. Hakenbeck, (1992). Relatedness of penicillin-binding protein 1A gene from different clones of penicillin resistant *Streptococcus pneumoniae* isolated in South Africa and Spain. EMBO Journal **11:** 3831-3836.

Martin, R., (1994). On the relationship between preferred termination codon contexts and nonsense suppression in human cells. Nucleic Acids Research **22:** 15-19.

Martin, S. L., B. Vrhovski and A. S. Weiss, (1995). Total synthesis and expression in *Escherichia coli* of a gene encoding human tropoelastin. Gene **154:** 159-166.

Martinussen, J., and K. Hammer, (1994). Cloning and characterization of *upp*, a gene encoding uracil phosphoribosyltransferase from *Lactococcus lactis*. Journal of Bacteriology **176:** 6457-6463.

Maruyama, T., T. Gojobori, S. Aota and T. Ikemura, (1986). Codon usage tabulated from the GenBank genetic sequence data. Nucleic Acids Research **14:** R151-R197.

Matassi, G., L. M. Montero, J. Salinas and G. Bernardi, (1989). The isochore organization and the compositional distribution of homologous coding sequences in the nuclear genome of plants. Nucleic Acids Research **17:** 5273-5290.

Matassi, G., P. M. Sharp and C. Gautier, (1999). Chromosomal location effects on gene sequence evolution in mammals. Current Biology **9:** 786-791.

Matic, I., M. Radman and C. Rayssiguier, (1994). Structure of recombinants from conjugational crosses between *Escherichia coli* donor and mismatch repair deficient *Salmonella typhimurium*. Genetics **136:** 16-26.

Matic, I., C. Rayssiguier and M. Radman, (1995). Interspecies gene exchange in bacteria: the role of SOS and mismatch repair systems in evolution of species. Cell **80:** 507-515.

Maynard Smith, J., and N. H. Smith, (1986). Site specific codon bias in bacteria. Genetics **142:** 1037-1043.

McInerney, J. O., (1997). Prokaryotic genome evolution as assessed by multivariate analysis of codon usage patterns. Microbiology Computational Genomics **2:** 1-10.

McInerney, J. O., (1998). Replicational and transcriptional selection on codon usage in *Borrelia burgdorferi*. Proceedings of The National Academy of Sciences of The United States of America **95:** 10698-10703.

McNally, T., I. Purvis, L. Fothergill-Gilmore and A. Brown, (1989). The yeast pyruvate kinase gene does not contain a string of non-preferred codons: revised nucleotide sequence. FEBS Letters **247:** 312-316.

Medigue, C., T. Rouxel, P. Vigier, A. Henaut and A. Danchin, (1991). Evidence for horizontal gene transfer in *Escherichia coli* speciation. Journal of Molecular Biology **222:** 851-856.

Merino, E., P. Balbas, J. L. Puente and F. Bolivar, (1994). Antisense overlapping open reading frames in genes from bacteria to humans. Nucleic Acids Research **22:** 1903-1908.

Milhon, J. L., and J. W. Tracy, (1995). Updated codon usage in *Schistosoma*. Experimental Parasitology **80:** 353-356.

Modiano, G., G. Battistuzzi and A. Motulsky, (1981). Nonrandom pattern of codon usage and of nucleotide substitution in human alpha- and beta-globin genes: an evolutionary strategy reducing the rate of mutations with drastic effects. Proceedings of The National Academy of Sciences of The United States of America **78:** 1110-1114

Mohsen, A. W. A., and J. Vockley, (1995). High-level expression of an altered cDNA-encoding human isovaleryl-CoA dehydrogenase in *Escherichia coli*. Gene **160:** 263-267.

Moriyama, E. N., and T. Gojobori, (1992). Rates of synonymous substitution and base composition of nuclear genes in *Drosophila*. Genetics **130:** 855-864.

Moriyama, E. N., and D. L. Hartl, (1993). Codon usage bias and base composition of nuclear genes in *Drosophila*. Genetics **134:** 847-858.

Morrison, D. A., J. Ellis and A. M. Johnson, (1994). An empirical comparison of distance matrix techniques for estimating codon usage divergence. Journal of Molecular Evolution **39:** 533-536.

Morton, B. R., (1994). Codon use and the rate of divergence of land plant chloroplast genes. Molecular Biology and Evolution **11:** 231-238.

Moszer, I., P. Glaser and A. Danchin, (1995). SubtiList: a relational data base for the *Bacillus subtilis* genome. Microbiology **141:** 261-268.

Mouchiroud, D., and C. Gautier, (1990). Codon usage changes and sequence dissimilarity between human and rat. Journal of Molecular Evolution **31:** 81-91.

Mrazek, J., and S. Karlin, (1998). Strand compositional asymmetry in bacterial and large viral genomes. Proceedings of The National Academy of Sciences of The United States of America **95:** 3720-3725.

Murray, E., J. Lotzer and M. Eberle, (1989). Codon usage in plant genes. Nucleic Acids Research **17:** 477-498.

Musto, H., H. Rodriguez Maseda and G. Bernardi, (1994). The nuclear genomes of African and American trypanosomes are strikingly different. Gene **141:** 63-69.

Muto, A., Y. Kawauchi, F. Yamo and S. Osawa, (1984). Preferential use of A- or U-rich codons for *Mycoplasma capricolum*. Nucleic Acids Research **12:** 8209-8217.

Muto, A., F. Yamao, Y. Kawauchi and S. Osawa, (1985). Codon usage in *Mycoplasma capricolum*. Proceedings of The Japan Academy Series B-Physical and Biological Sciences **61:** 12-15.

Nakamura, Y., K. Wada, Y. Wada, H. Doi, S. Kanaya *et al.*, (1996). Codon usage tabulated from the international DNA sequence databases. Nucleic Acids Research **24:** 214-215.

Nassal, M., T. Mogi, S. S. Karnik and H. G. Khorana, (1987). Structure-function studies on bacteriorhodopsin. III. Total synthesis of a gene for bacterio-opsin and its expression in *Escherichia coli*. Journal of Biological Chemistry **262:** 9264-9270.

Nei, M., and D. Graur, (1984). Extent of protein polymorphism and the neutral theory. Evolutionary Biology **17:** 73-118.

Nesti, C., G. Poli, M. Chicca, P. Ambrosino, C. Scapoli *et al.*, (1995). Phylogeny inferred from codon usage pattern in 31 organisms. Computer Applications for the Biosciences **2:** 167-171.

Newgard, C. B., K. Nakano, P. K. Hwang and R. J. Fletterick, (1986). Sequence analysis of the cDNA encoding human liver glycogen phosphorylase reveals tissue specific codon usage. Proceedings of The National Academy of Sciences of The United States of America **83:** 8132-8136.

Nichols, B., G. Miozzari, M. van Cleemput, G. Bennett and C. Yanofsky, (1980). Nucleotide sequence of the *trpG* regions of *Escherichia coli, Shigella dysenteriae, Salmonella typhimurium* and *Serratia marcescens*. Journal of Molecular Biology **142:** 503-517.

Nilsson, L., and V. Emilsson, (1994). Factor for inversion stimulation dependent growth rate regulation of individual transfer-RNA species in *Escherichia coli*. Journal of Biological Chemistry, **269:** 9460- 9465.

Noltmann, E. A., (1972). *Aldose-ketose isomerase. Triose-phosphate isomerase.* Academic Press, London.

Nomura, M., F. Sor, M. Yamagashi and M. Lawson, (1987). Heterogeneity of GC content within a single bacterium and its implications for evolution. Cold Spring Harbor Symposium Quantitative Biology .

Nussinov, R., (1981). Eukaryotic dinucleotide preference rules and their implications for degenerate codon usage. Journal of Molecular Biology **149:** 125-131.

Nussinov, R., (1984). Strong doublet preferences in nucleotide sequences and DNA geometry. Journal of Molecular Evolution **20:** 111-119.

Ochman, H., and A. Wilson, (1987a). Evolutionary history of enteric bacteria., pp. 1649 in *Escherichia coli* and *Salmonella typhimurium*, edited by Neidhardt. ASM Press, Washington DC.

Ochman, H., and A. C. Wilson, (1987b). Evolution in Bacteria: evidence for a universal substitution rate in cellular genomes. Journal of Molecular Evolution **26:** 74-86.

Ogasawara, N., (1985). Markedly unbiased codon usage in *Bacillus subtilis*. Gene **40:** 145-150.

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

Ohno, S., (1988). Codon preference is but an illusion created by the construction principle of coding sequences. Proceedings of The National Academy of Sciences of The United States of America **85:** 4378-4382.

Oliver, S. G., Q. J. M. Vanderaart, M. L. Agostonicarbone, M. Aigle, L. Alberghina *et al.*, (1992). The complete DNA sequence of yeast Chromosome III. Nature, **357:** 38-46.

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

Osawa, S., and T. H. Jukes, (1989). Codon reassignment (codon capture) in evolution. Journal of Molecular Evolution **29:** 271-278.

Osawa, S., T. H. Jukes, K. Watanabe and A. Muto, (1992). Recent evidence for evolution of the genetic code. Microbiology Reviews **56:** 229-264.

Osawa, S., A. Muto, T. Jukes and T. Ohama, (1990). Evolutionary changes in the genetic code. Proceedings of the Royal Society of London Series B **241:** 19-28.

Osawa, S., T. Ohama, F. Yamao, A. Muto, T. Jukes *et al.*, (1988). Directional mutation pressure and transfer RNA in choice of the third nucleotide of synonymous two codon sets. Proceedings of The National Academy of Sciences of The United States of America **85:** 1124-1128.

Oskouian, B., and G.C. Stewart, (1990). Repression and catabolite repression of the lactose operon of *Staphylococcus aureus*. Journal of Bacteriology **172**: 3804-3812

Ossadnik, S. M., S. V. Buldyrev, A. L. Goldberger, S. Havlin, R. N. Mantegna *et al.*, (1994). Correlation approach to identify coding regions in DNA sequences. Biophysical Journal **67:** 64-70.

Pakrasi, H. B., (1995). Genetic analysis of the form and function of photosystem I and photosystem II. Annual Review of Genetics **29:** 755-776.

Parker, J., T. C. Johnston, P. T. Borgia, G. Holtz, E. Remaut *et al.*, (1983). Codon usage and mistranslation - *in vivo* basal level misreading of the MS2 coat protein message. Journal of Biological Chemistry **258:** 7-12.

Parker, J., J. Precup and C. W. Fu, (1992). Misreading of the *argl* message in *Escherichia coli*. FEMS Microbiology Letters **100:** 141-145.

Pedersen, S., P. L. Bloch, S. Keeh and F. C. Neidhardt, (1978). Patterns of protein synthesis in *E. coli*: A catalogue of the amount of 140 individual proteins at different growth rates. Cell **14:** 179-190.

Precup, J., and J. Parker, (1987). Missense misreading of Asparagine codons as a function of codon identity and context. Journal of Biological Chemistry **262:** 11351-11356.

Perriere, G., and M. Gouy, (1996). WWW-Query an online sequence retrieval system for biological sequence banks. Biochimie **78:** 364-369.

Perriere, G., M. Gouy and T. Gojobori, (1994). NRSUB a nonredundant database for the *Bacillus subtilis* genome. Nucleic Acids Research **22:** 5525-5529.

Perriere, G., and J. Thioulouse, (1996). Online tools for sequence retrieval and multivariate-statistics in molecular-biology. Computer Applications In The Biosciences, **12:** 63-69.

Petersen, C., (1987). The functional stability of the lacZ transcript is sensitive towards sequence alterations immediately downstream of the ribosome binding site. Molecular and General Genetics **209:** 179-187.

Petersen, G. B., P. A. Stockwell and D. F. Hill, (1988). Messenger RNA recognition in *Escherichia coli* - a possible second site of interaction with 16S ribosomal RNA. EMBO journal **7:** 3957-3962.

Pfitzinger, H., P. Guillemaut, J. H. Weil and D. T. N. Pillay, (1987). Adjustment of the tRNA population to the codon usage of chloroplasts. Nucleic Acids Research **15:** 1377.

Phillips, G. J., J. Arnold and R. Ivarie, (1987a). The effect of codon usage on the oligonucleotide composition of the *Escherichia coli* genome and identification of overrepresented and underrepresented sequences by Markov chain analysis. Nucleic Acids Research **15:** 2627-2638.

Phillips, G. J., J. Arnold and R. Ivarie, (1987b). Mono through hexanucleotide composition of the *Escherichia coli* genome: a Markov chain analysis. Nucleic Acids Research **15:** 2611-2626.

Poole, E. S., C. M. Brown and W. P. Tate, (1995). The identity of the base following the stop codon determines the efficiency of *in vivo* translational termination in *Escherichia coli*. EMBO Journal **14:** 151-158.

Poolman, B., (1993). Energy transduction in lactic-acid bacteria. FEMS Microbiology Reviews **12:** 125-148.

Popplewell, A. G., M. G. Gore, M. Scawen and T. Atkinson, (1991). Synthesis and mutagenesis of an IgG-binding protein based upon protein A of *Staphylococcus aureus*. Protein Engineering **4:** 963-970.

Porter, E. V., and B. M. Chassy, (1988). Nucleotide sequence of the β–D-phosphogalactosidase galactohydrolase gene of *Lactobacillus casei*: comparison to analogous *pbg* genes in other Gram-positive species. Gene **62:** 263-276.

Post, L., G. Strycharz, M. Norma, H. Lewis and P. Dennis, (1979). Nucleotide sequence of the ribosomal protein gene cluster adjacent to the gene for RNA polymerase subunit in *Escherichia coli*. Proceedings of The National Academy of Sciences of The United States of America **76:** 1697-1701.

Pouwels, P. H., and J. A. M. Leunissen, (1994). Divergence in codon usage of *Lactobacillus* species. Nucleic Acids Research **22:** 929-936.

Prevots, F., E. Remy, M. Mata and P. Ritzenthaler, (1994). Isolation and characterization of large lactococcal phage resistance plasmids by pulsed-field gel-electrophoresis. FEMS Microbiology Letters **117:** 7-13.

Purvis, I.J., A. J.E. Bettany, T.C. Santiago, J.R. Coggins, K. Duncan *et al.*, (1987). The efficiency of folding of some proteins is increased by controlled rates of translation *in vivo* - a hypothesis. Journal of Molecular Biology **193:** 413-417.

Reeves, P., (1993). Evolution of *Salmonella* O antigen by interspecific gene transfer on a large scale. Trends in Genetics **9:** 17-22.

Rex, G., B. Surin, G. Besse, B. Schneppe and J. E. G. McCarthy, (1994). The mechanism of translational coupling in *Escherichia coli* - higher-order structure in the *atpHA* messenger-RNA acts as a conformational switch regulating the access of *de novo* initiating ribosomes. Journal of Biological Chemistry **269:** 18118-18127.

Robinson, M., R. Lilley, S. Little, J. S. Emtage, G. Yarranton *et al.*, (1984). Codon usage can affect efficiency of translation of genes in *Escherichia coli*. Nucleic Acids Research **12:** 6663-6671.

Rodriguez Belmonte, E., M. A. Freire Picos, A. M. Rodriguez Torres, M. I. Gonzalez Siso, M. E. Cerdan *et al.*, (1996). PICDI, a simple program for codon bias calculation. Molecular Biotechnology, **5:** 191-195.

Romero, D. A., and T. R. Klaenhammer, (1993). Transposable elements in lactococci - a review. Journal of Dairy Science **76:** 1-19.

Rosenberg, A. H., E. Goldman, J. J. Dunn, F. W. Studier and G. Zubay, (1993). Effects of consecutive AGG codons on translation in *Escherichia coli*, demonstrated with a versatile codon test system. Journal of Bacteriology **175:** 716-722.

Rosey, E., and G. Stewart, (1989). The nucleotide sequence of the *lacC* and *lacD* genes of *Staphylococcus aureus*. Nucleic Acids Research **17:** 3980.

Rosey, E. L., B. Oskouian and G. C. Stewart, (1991). Lactose metabolism by *Staphylococcus aureus* - characterization of lacABCD, the structural genes of the tagatose 6-phosphate pathway. Journal of Bacteriology **173:** 5992-5998.

Rosey, E. L., and G. C. Stewart, (1993). Nucleotide and deduced amino acid sequences of the *lacR,lacABCD*, and *lacFE* genes encoding the repressor, tagatose-6-phosphate gene cluster, and sugar specific phosphotransferase system components of the lactose operon of *Streptococcus mutans.* Journal of Bacteriology **174:** 6159-6170.

Rubin, G. M., (1998). The Drosophila genome project: a progress report. Trends In Genetics, **14:** 340-343.

Sabeur, G., G. Macaya, F. Kadi and G. Bernardi, (1993). The isochore patterns of mammalian genomes and their phylogenetic implications. Journal of Molecular Evolution **37:** 93-108.

Sahl, H. G., (1994). Gene-encoded antibiotics made in bacteria. Ciba Foundation Symposia **186:** 27-42.

Saier, M. J., (1995). Differential codon usage: a safe guard against inappropriate gene expression of specialized genes. FEBS **362:** 1-4.

Sanger, F., G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson *et al.*, (1977). Nucleotide sequence of bacteriophage ΦX174. Nature **265:** 687-695.

Sayers, J. R., H. P. Price, P. G. Fallon and M. J. Doenhoff, (1995). AGA/AGG codon usage in parasites - implications for gene expression in *Escherichia coli*. Parasitology Today **11:** 345-346.

Scherer, S., M. S. McPeek and T. P. Speed, (1994). Atypical regions in large genomic DNA sequences. Proceedings of the National Academy of Sciences of the United States of America **91:** 7134-7138.

Schleifer, K. H., J. Kraus, C. Dvorak, R. Kilpperbalz, M. D. Collins *et al.*, (1985). Transfer of *Streptococcus lactis* and related *Streptococci* to the genus *Lactococcus* gen-nov. Systematic and Applied Microbiology **6:** 183-195.

Schmidt, J., M. Bubunenko and A. R. Subramanian, (1993). A novel operon organisation involving the genes for chorismate synthase (aromatic biosynthesis pathway) and ribosomal GTPase center proteins (L11, L1, L10, L12: *rplKAJL*) in cyanobacterium *Synechocystis* PCC 6803. Journal of Biological Chemistry **268:** 27447-27457.

Schmidt, W., (1995). Phylogeny reconstruction for protein sequences based on amino acid properties. Journal of Molecular Evolution **41:** 522-530.

Sedgwick, S., M. Robson and F. Malik, (1988). Polymorphisms in the *umuCD* region of *Escherichia* species. Journal of Bacteriology **170:** 1610-1616.

Selby, C., and A. Sancar, (1993). Transcriptional repair coupling and mutation frequency decline. Journal of Bacteriology **175:** 7509-7514.

Shanley, M. S., A. Harrison, R. E. Parales, G. Kowalchuck, D. Mitchell *et al.*, (1994). Unusual G+C content and codon usage in *catIJF*, a segment of the *ben cat supra* operonic cluster in the *Acinetobacter calcoaceticus* chromosome. Gene **138:** 59-65.

Sharp, P., (1985). Does the 'non-coding' strand code. Nucleic Acids Research **13:** 1389-1397.

Sharp, P., N. Nolan and K. Devine, (1995a). Evolution of gene sequences between and within species of *Bacillus*, pp. in *Populaton genetics of bacteria*, edited by S. Baumberg, J. Young, S. Saunders and E. Wellington. Society for General Microbiology, London.

Sharp, P. M., (1986). What can aids virus codon usage tell us. Nature **324:** 114.

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

Sharp, P. M., (1991). Determinants of DNA sequence divergence between *Escherichia coli* and *Salmonella typhimurium* codon usage, map position, and concerted evolution. Journal of Molecular Evolution **33:** 23-33.

Sharp, P. M., M. Averof, A. T. Lloyd, G. Matassi and J. F. Peden, (1995b). DNA sequence evolution - the sounds of silence. Philosophical Transactions of the Royal Society of London Series B-biological Sciences **349:** 241- 247.

Sharp, P. M., and M. Bulmer, (1988). Selective differences among translation termination codons. Gene **63:** 141-145.

Sharp, P. M., C. J. Burgess, A. T. Lloyd and K. J. Mitchell, (1992).*Selective use of termination codons and variations in codon choice.* CRC press, London.

Sharp, P. M., and E. Cowe, (1991). Synonymous codon usage in *Saccharomyces cerevisiae*. Yeast **7:** 657-678.

Sharp, P. M., E. Cowe, D. G. Higgins, D. C. Shields, K. H. Wolfe *et al.*, (1988). Codon usage patterns in *Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster* and *Homo sapiens* - a review of the considerable within species diversity. Nucleic Acids Research **16:** 8207-8211.

Sharp, P. M., and K. M. Devine, (1989). Codon usage and gene-expression level in *Dictyostelium discoideum* - highly expressed genes do prefer optimal codons. Nucleic Acids Research **17:** 5029-5039.

Sharp, P. M., D. G. Higgins, D. C. Shields, K. M. Devine and J. A. Hoch, (1990). *Bacillus subtilis gene sequences*. Academic Press, Orlando.

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

Sharp, P. M., and W. H. Li, (1987a). The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Research **15:** 1281-1295.

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

Sharp, P. M., and W. H. Li, (1989). On the rate of DNA-sequence - evolution in *Drosophila*. Journal of Molecular Evolution **28:** 398-402.

Sharp, P. M., and A. T. Lloyd, (1993). Regional base composition variation along yeast chromosome III evolution of chromosome primary structure. Nucleic Acids Research **21:** 179-183.

Sharp, P. M., and G. Matassi, (1994). Codon usage and genome evolution. Current Opinions in Genetics and Development **4:** 851-860.

Sharp, P. M., D. C. Shields, K. H. Wolfe and W. H. Li, (1989). Chromosomal location and evolutionary rate variation in Enterobacterial genes. Science **246:** 808-810.

Sharp, P. M., M. Stenico, J. F. Peden and A. T. Lloyd, (1993). Codon usage - mutational bias, translational selection, or both. Biochemical Society Transactions **21:** 835-841.

Sharp, P. M., T. M. F. Tuohy and K. R. Mosurski, (1986). Codon usage in yeast cluster-analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Research **14:** 5125-5143.

Sharp, P. M., and F. Wright, (1988). Analysis of yeast DNA sequence data: codon usage of the distantly related yeasts *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. Yeast **4:** S515.

Shields, D., (1989). Evolution of codon usage patterns, in *Department of Genetics*. Trinity College Dublin, Dublin.

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

Shields, D. C., and P. M. Sharp, (1987). Synonymous codon usage in *Bacillus subtilis* reflects both translational selection and mutational biases. Nucleic Acids Research **15:** 8023-8040.

Shields, D. C., P. M. Sharp, D. G. Higgins and F. Wright, (1988). Silent sites in *Drosophila* genes are not neutral - evidence of selection among synonymous codons. Molecular Biology and Evolution **5:** 704-716.

Shpaer, E., (1986). Constraints on codon context in *Escherichia coli*. Journal Molecular Biology **188:** 555-564.

Simon, M., J. Zieg, M. Silverman, G. Mandel and R. Doolittle, (1980). Phase variation: evolution of a controlling element. Science **209:** 1370-1374.

Smith, D. R., L. A. Doucettestamm, C. Deloughery, H. M. Lee, J. Dubois *et al.*, (1997). Complete genome sequence of *Methanobacterium thermoautotrophicum* delta h: functional analysis and comparative genomics. Journal of Bacteriology **179:** 7135-7155.

Smith, D. W. E., (1975). Reticulocyte transfer RNA and haemoglobin synthesis. Science **190:** 529-535.

Smith, N., P. Beltran and R. Selander, (1990). Recombination of *Salmonella* phase 1 flagellum genes generates new serovars. Journal of Bacteriology **172:** 2209-2216.

Smith, T.F., W.W. Ralph, M. Goodman and J. Czelusniak, (1985). Codon usage in vertebrate hemoglobins and its implications. Molecular Biology and Evolution **2:** 390-398.

Sneath P.H., (1986). Streptococcus, pp. 1059-1061 in *Bergey's Manual of Systematic Bacteriology*, edited by P. H. Sneath, N. S. Mair, M. E. Sharpe and J. G. Holt. Williams and Wilkins, London.

Soltesrak, E., D. J. Kushner, D. D. Williams and J. R. Coleman, (1995). Factors regulating *cryivb* expression in the Cyanobacterium *Synechococcus*PCC-7942. Molecular & General Genetics **246:** 301-308.

Sorensen, M. A., C. G. Kurland and S. Pedersen, (1989). Codon usage determines translation rate in *Escherichia coli*. Journal of Molecular Biology **207:** 365-377.

Spanjaard, R. A., and J. van Duin, (1988). Translation of the sequence AGG-AGG yields 50% ribosomal frame shifting. Proceedings of the National Academy of Sciences of the United States of America **85:** 7967-7971.

Spratt, B. G., L. D. Bowler, J. Zhang and J. Maynard-Smith, (1992). Role of interspecies transfer of chromosomal genes in the evolution of penicillin resistance in pathogenic and commensal *Neisseria* species. Journal of Molecular Evolution **34:** 115-125.

Sprengart, M. L., H. P. Falscher and E. Fuchs, (1990). The initiation of translation in *E. coli*: apparent base pairing between the 16S rRNA and downstream sequences of the mRNA. Nucleic Acids Research **18:** 1719-1723.

Staden, R., and A. D. Mclachlan, (1982). Codon preference and its use in identifying protein coding regions in long DNA sequences. Nucleic Acids Research **10:** 141-156.

Stenico, M., A. T. Lloyd and P. M. Sharp, (1994). Codon usage in *Caenorhabditis elegans* - delineation of translational selection and mutational biases. Nucleic Acids Research **22:** 2437-2446.

Sueoka, N., (1961). Variation and heterogeneity of base composition compilation of old and new data. Journal of Molecular Biology **3:** 31-40.

Sueoka, N., (1962). On the genetic basis of variation and heterogeneity of DNA base composition. Proceedings of the National Academy of Sciences of the United States of America **48:** 582-592.

Sueoka, N., (1988). Directional mutational pressure and neutral pressure. Proceedings of the National Academy of Sciences of the United States of America **85:** 2653-2657.

Sueoka, N., (1992). Directional mutation pressure, selective constraints, and genetic equilibria. Journal of Molecular Evolution **34:** 95-114.

Suzuki, Y., and D. Brown, (1972). Isolation and identification of the messenger RNA for silk fibroin from *Bombyx mori*. Journal of Molecular Biology **63:** 409-429.

Tandeau de Marsac, N., and J. Houmard, (1987). *Advances in Cyanobacterial Molecular Genetics*. Elsevier.

Tate, W., (1984).*Termination of protein synthesis*. Marcel Dekker, New York.

Thioulouse, J., (1990a). MacMul and GraphMul: two Macintosh programmes for the display and analysis of multivariate data. Computers and Geosciences **16:** 1235-1240.

Thioulouse, J., (1990b). Statistical analysis and graphical display of multivariate data on the Macintosh. Computer Applications in the Biosciences **5:** 287-292.

Thioulouse, J., (1996). Towards better graphics for multivariate-analysis - the interactive factor map. Computational Statistics, **11:** 11-21.

Thioulouse, J., and F. Chevenet, (1996). Netmul, a world-wide-web user interface for multivariate analysis software. Computational Statistics & Data Analysis, **21:** 369-372.

Thioulouse, J., S. Doledec, D. Chessel and J. M. Oliver, (1995). ADE softeware: multivariate analysis and graphical display of environmental data, pp. 57-61 in *Sofware per l'Ambiente*, edited by G. Guariso, and A. Rizzoli. Patron editor, Bolonia.

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

Thompson, K., (1987). Regulation fo sugar transport and metabolism in lactic acid bacteria. FEMS Microbiological Reviews **46:** 221-231.

Tobais, J., T. Shrader, G. Rocap and A. Varshavsky, (1991). The N-end rule in bacteria. Science **259:** 1374-1377.

Tomb, J. F., O. White, A. R. Kerlavage, R. A. Clayton, G. G. Sutton *et al.*, (1997). The complete genome sequence of the gastric pathogen *Helicobacter pylori*. Nature **388:** 539-547.

Trifonov, E. N., (1987). Translation framing code and frame monitoring as suggested by the analysis of mRNA and 16S rRNA nucleotide sequences. Journal of Molecular Biology **194:** 643-652.

Tyson, H., and R. Dhindsa, (1995). Codon usage in plant peroxidase genes. DNA Sequence **5:** 339-351.

Ueda, Y., S. Taguchi, K. Nishiyama, I. Kumagai and K. Miura, (1993). Effect of rare leucine codon, TTA on expression of a foreign gene in *Streptomyces lividans*. Biochimica et Biophysica Acta **1172:** 261-266.

van de Guchte, M., J. Kok and G. Venema, (1992). Gene-expression in *Lactococcus lactis*. FEMS Microbiology Reviews **88:** 73-92.

van de Guchte, M., T. van der Lende, J. Kok and G. Venema, (1991). A possible contribution of messenger-RNA secondary structure to translation initiation efficiency in *Lactococcus lactis*. FEMS Microbiology Letters **81:** 201-208.

van Rooijen, R. J., S. van Schalkwijk and W. M. Devos, (1991). Molecular-cloning, characterization, and nucleotide-sequence of the tagatose 6-phosphate pathway gene-cluster of the lactose operon of *Lactococcus lactis*. Journal of Biological Chemistry **266:** 7176-7181.

Varenne, S., J. Buc, R. Llouss and C. Lazdubnski, (1984). Translation of a non-uniform process: effect of tRNA availability on the rate of elongation of nascent polypeptide chains. Journal of Molecular Biology **180:** 549-576.

Verma, N., and P. Reeves, (1989). Identification and sequences of *rbfS* and *rbfE*, which determine antigenic specificity of group A and group D *Salmonellae*. Journal of Bacteriology **171:** 5694-5701.

Wada, K., Y. Wada, H. Doi, F. Ishibashi, T. Gojobori *et al.*, (1991). Codon usage tabulated from the GenBank genetic sequence data. Nucleic Acids Research **19:** 1981-1986.

Wada, K. N., Y. Wada, F. Ishibashi, T. Gojobori and T. Ikemura, (1992). Codon usage tabulated from the GenBank genetic sequence data. Nucleic Acids Research **20:** 2111-2118.

Wadman, M., (1999). Human Genome Project aims to finish 'working draft' next year. Nature **398:** 177.

Wainer, H., (1983). On Multivariate display, pp. 469-508 in *Recent Advances in Statistics*, edited by M. H. Rizzi, J. S. Rustagi and D. Siegmund. Academic Press, New York.

Wang, B. Q., L. Lei and Z. Burton, (1994). Importance of codon preference for production of human RAP74 and reconstitution of the RAP30/74 complex. Protein Engineering Purification **5:** 476-485.

Weiss, R. B., D. M. Dunn, J. F. Atkins and R. F. Gesteland, (1987). Slippery runs, shifty stops, backward steps, and forward hops -2, -1, +1, +2, +5, and +6 ribosomal frameshifting. Cold Spring Harbor Symposia On Quantitative Biology **52:** 687-693.

Weiss, R. B., D. M. Dunn, A. E. Dahlberg, J. F. Atkins and R. F. Gesteland, (1988). Reading frame switch caused by base pair formation between the 3' end of 16S ribosomal-RNA and the messenger-RNA during elongation of protein synthesis in *Escherichia coli*. EMBO Journal **7:** 1503-1507.

Wikstrom, P. M., L. K. Lind, D. E. Berg and G. R. Bjork, (1992). Importance of messenger-RNA folding and start codon accessibility in the expression of genes in a ribosomal protein operon of *Escherichia coli*. Journal of Molecular Biology **224:** 949-966.

Wilmotte, A., (1994). Molecular evolution of the cyanobacteria, pp. 1-25 in *Molecular Biology of Cyanobacteria*, edited by D. Bryant. Kluwer, Dordrech.

Winkler, H., and D. Wood, (1988). Codon usage in selected AT-rich bacteria. Biochimie **70:** 977-986.

Woese, C. R., (1987). Bacterial evolution. Microbiology Reviews **51:** 227-271.

Wolfe, K. H., P. M. Sharp and W. H. Li, (1989). Mutation-rates differ among regions of the mammalian genome. Nature **337:** 283-285.

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

Wright, F., and M. J. Bibb, (1992). Codon usage in the G+C rich *Streptomyces* genome. Gene **113:** 55-65.

Wu, L.-F., and M. J. Saier, (1991). Differences in codon usage among genes encoding proteins of different function in *Rhodobacter capsulatus*. Research Microbiology **142:** 943-949.

Yarus, M., and L. Folley, (1985). Sense codons are found in specific contexts. Journal Molecular Biology **182:** 529-540.

Zhang, C. T., and K. C. Chou, (1994). A graphic approach to analyzing codon usage in 1562 *Escherichia coli* protein coding sequences. Journal of Molecular Biology **238:** 1-8.

Zhang, S., G. Zubay and E. Goldman, (1991). Low usage codons in *Escherichia coli*, yeast, fruit fly and primates. Gene **105:** 61-72.

f