A novel approach for predicting the evolutionary origin of acquisition of foreign genes in bacteria: Application of codon usage analyses

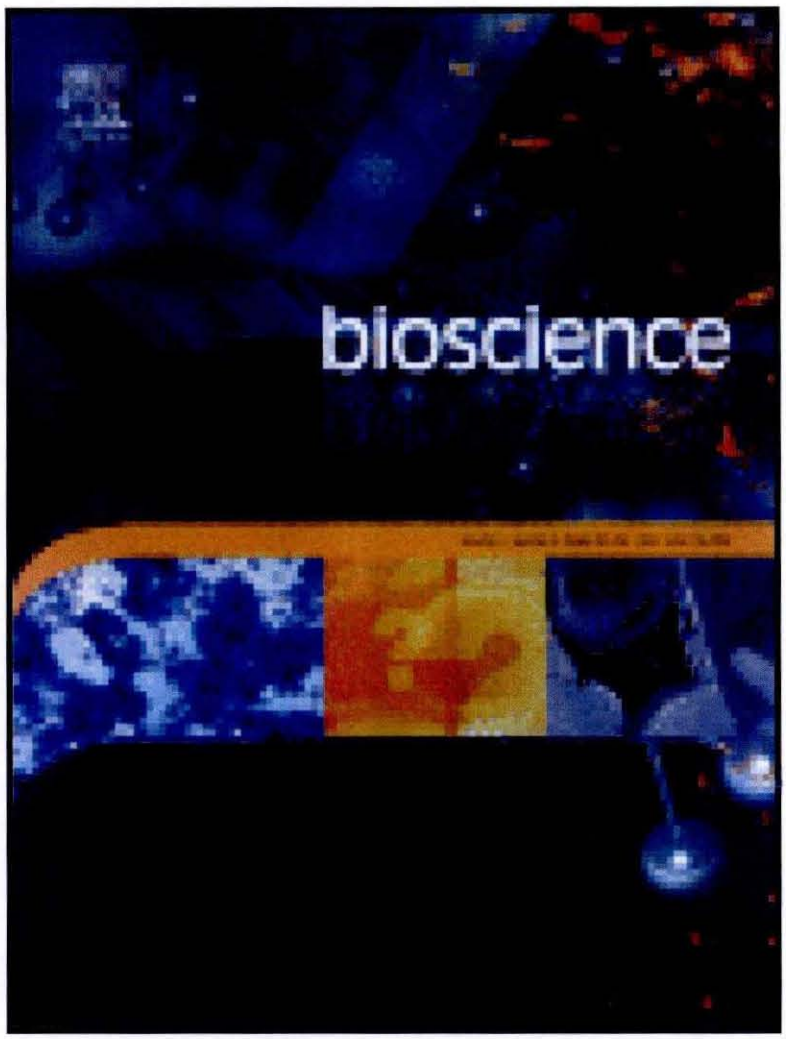
Submitted by

Md. Moazzem Hossain Khan

ID: 07276001
MS in Biotechnology
Department of Mathematics
& Natural Sciences
BRAC University

Thesis submitted to the BRAC University for the Degree of MS in Biotechnology

This article has been published in **Bioscience Hypotheses,**
Volume 2, Issue 4, 2009, Pages 217-222



Bioscience Hypotheses
Volume 2, Issue 4, 2009, Pages 217-222

Font Size

Abstract | Article | Figures/Tables | References | Purchase PDF (267 K)

doi:10.1016/j.bihy.2008.12.003
Cite or Link Using DOI

A novel approach for predicting the evolutionary time of acquisition of foreign genes by bacteria: Application of codon usage analyses

Sajib Chakraborty[a], Mohammad Arif[a], T.M. Zaved Waise[a], Farhana Naznin[a], Faizule Hassan[a], Mohammad Faisal[a], Md. Fakhrul Kabir[a], Md. Moazzem Hossain Khan[b], Md. Ehsanul Hoque Mazumder[c], Yearul Kabir[d] and Mark A. Smith[e]

[a]Department of Biochemistry and Molecular Biology, University of Dhaka, Dhaka 1000, Bangladesh

[b]Department of Biotechnology, BRAC University, Dhaka, Bangladesh

[c]Department of Pharmacy, Jahangirnagar University, Savar, Bangladesh

[d]Department of Family Sciences, College for Women, Kuwait University, Kuwait City, Kuwait

[e]Department of Pathology, Case Western Reserve University, Cleveland, OH, USA

Received 26 November 2008; revised 20 December 2008; accepted 28 December 2008. Available online 9 May 2009.

Abstract

Table of Contents

ABSTRACT

Lysogenic bacteriophages, are considered as a major player for the introduction of foreign genes into bacterial strains. At the time of introduction foreign genes do not fit well into the translation system of the recipient host bacterium as they tend to retain the characteristics of the donor bacteriophages from which it has been transferred. Consequently foreign genes are poorly transcribed at the early phase of their evolution within the host bacterium. This is largely due to the difference in the codon usage pattern between the horizontally transferred genes and the host bacterium. In this study we present detail analyses of various parameters of the codon usages such as codon adaptation index (CAI), mean difference (MD) of the relative adaptiveness, synonymous substitution rate (SSR) of six different phage encoded toxin genes (cholera toxin, shiga toxin, diphtheria toxin, neurotoxin C1, enterotoxin type A and cytotoxin) and proposed conceptual relationship between the evolutionary time of acquisition of the foreign genes and the selected set of parameters of the codon usage. On the basis of the observed data we hypothesize that CAI, MD and SSR of the phage encoded toxin genes are correlated with the evolutionary time of their acquisition and developed a novel approach based on the analyses of these parameters, that can be used to predict the evolutionary time of their acquisition by the corresponding host bacterium.

ACKNOWLEDGEMENT

This is a golden opportunity for me to convey my sincere regards for all those people who enabled me to accomplish my dissertation work successfully. It is my privilege to express my sincere gratitude to my supervisor **Dr. Md. Anwar Hossain**, Department of Biochemistry and Molecular Biology, University of Dhaka, Bangladesh, for his guidance, encouragement and valuable advice. It is his confidence imbibing attitude and splendid discussions and endless endeavors through which I have gained significant experience. My sincere thanks to Dr. Mohammed Arif, Department of Biochemistry and Molecular Biology, University of Dhaka, Bangladesh, for his immense concern throughout the project work. And my special thanks to Prof. Naiyyum Choudhury, Co-ordinator, Biotechnology program, Department of Mathematics & Natural Sciences, BRAC University for his favor and concern.

I am thankful for the help rendered by Mr. Sajib Chakraborty, and Mr. T.M. Zaved Waise throughout my project period without which it would have been difficult for me to achieve my goals.

Special thanks to all faculty members and staff, BRAC University and Dhaka University, for their constant encouragement and support throughout the project work. I am thankful for the help rendered by, who were always there for helping me to understand the minor details of the work on a day-to-day basis.

I feel deep gratitude to Dr. Zesmin, Chairman, Department of Genetic Engineering and Biotechnology, Dhaka University for giving me first hand learning of Bioinformatics during completion of Bioinformatics course.

I feel lacunae of words to express my most heartfelt and cordial thanks to my friends Farhana Nazin, Faizule Hassan, Mohammad Faisal, Md. Fakrul Kabir, Md. Ehsanul Hoque Mazumder, Yearul Kabir who have always stood by my side during all the tough times.

The whole credit of my achievements during the project work goes to my parents and family members and also my employer i.e; SQUARE Pharmaceuticals Ltd. It was their unshakeable faith in me that has always helped me to proceed further.

I dedicate this thesis to those who believed in me, my parents Late Anwar Hossain Khan and Mrs. Ashura Khanam.

In the end, I am thankful to the Almighty Allah for blessing me to complete this work successfully.

Chapter 1

Introduction

Degeneracy of the genetic code allows synonymous codons to code for the same amino acid. In a particular species several synonymous codons are utilized more frequently than others during protein synthesis. The pattern of choices between synonymous codons varies from one gene to another according to the type of genome the gene occurs in. Thus codon usage is mainly a genome strategy, contrary to amino acid usage in proteins. This non randomness in the utilization of the synonymous codons is believed to be arisen from the mutational biases and various selective forces. It is argued that the bias in synonymous codon usage observed in unicellular organisms is due to a balance between the forces of selection and mutation in a finite population, with greater bias in highly expressed genes reflecting stronger selection for efficiency of translation. A population genetic model is developed taking into account population size and selective differences between synonymous codons. A biochemical model is then developed to predict the magnitude of selective differences between synonymous codons in unicellular organisms in which growth rate (or possibly growth yield) can be equated with fitness. Selection can arise from differences in either the speed or the accuracy of translation. A model for the effect of speed of translation on fitness is considered in detail, a similar model for accuracy more briefly. The model is successful in predicting a difference in the degree of bias at the beginning than in the rest of the gene under some circumstances, as observed in *Escherichia coli*, but grossly overestimates the amount of bias expected. G+C composition of the genome is a vital factor for codon usage variation. This variation mostly lies in the third position of the codons (<20% to >90% G+C), as it is immune to changes. GC-rich organisms tend to prefer GC-containing codons over AT-containing ones. Consequently each organism has their optimal and nonoptimal codons.

Bacteria can acquire foreign genes through HGT (horizontal gene transfer). Bacteriophages are the major player in the HGT phenomenon. Bacteriophage can mobilize genetic material between distantly related bacterial species. At the time of introduction into the recipient host bacterium, the foreign genes tend to retain the characteristics of the donor bacterium and it may vary significantly from the native genes of the recipient bacterium in terms of optimal codon usage. For the detection of the horizontally transferred foreign genes various parameters of the codon usage such as relative adaptiveness (RA), mean difference of RA, codon adaptation index (CAI), synonymous substitution rate (SSR) between codons can be used. CAI is a measure of similarity of a gene's synonymous codon usage to that of a standard set of highly expressed genes for that organism. The mean deference (MD) of the relative adaptiveness (RA) of the codons of the foreign genes from that of the native genes give us a clue about by what extent the foreign gene varies from the native genes in a host bacterium. Here we present a detailed analysis of a selected set of parameters such as RA, MD of RA, CAI, and SSR of the codon usage pattern of the six phage encoded toxin genes. These are cholera toxin, shiga toxin, neurotoxin C1, enterotoxins type A, cytotoxin and diphtheria toxin. To the best of our knowledge these parameters of the codon usage has not been utilized previously in predicting the time of acquisition of foreign genes. In previous method the rate of horizontal gene transfer was estimated, but not their evolutionary time of acquisition. In this study we proposed a hypothesis involving the conceptual relationship between the evolutionary time of acquisition of the foreign genes and the selected set of parameters of the codon usage and adopt a novel approach for the prediction of the comparative time of the acquisition of the foreign genes on the basis of the analyses the selected parameters.

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 øX174 (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 freeliving 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 *e 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 nonconserved residues. In regions encoding important protein motifs (homeodomains and zincfinger 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 sequencespecific 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 underrepresentation 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 "noncoding" 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 isochors (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 1st and 3rd 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 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

characterize 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 ($tRNA_{arg}$ 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 ($tRNA_{thr}$ 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 ($tRNA_{leu}$ 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 $tRNA_{leu}$ 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 $tRNA_{AGG}$ 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).

Chapter 2

Proposal

Degeneracy of the genetic code allows synonymous codons to code for the same amino acid. In a particular species several synonymous codons are utilized more frequently than others during protein synthesis. The pattern of choices between synonymous codons varies from one gene to another according to the type of genome the gene occurs in. Thus codon usage is mainly a genome strategy, contrary to amino acid usage in proteins. This non randomness in the utilization of the synonymous codons is believed to be arisen from the mutational biases and various selective forces. It is argued that the bias in synonymous codon usage observed in unicellular organisms is due to a balance between the forces of selection and mutation in a finite population, with greater bias in highly expressed genes reflecting stronger selection for efficiency of translation. A population genetic model is developed taking into account population size and selective differences between synonymous codons. A biochemical model is then developed to predict the magnitude of selective differences between synonymous codons in unicellular organisms in which growth rate (or possibly growth yield) can be equated with fitness. Selection can arise from differences in either the speed or the accuracy of translation. A model for the effect of speed of translation on fitness is considered in detail, a similar model for accuracy more briefly. The model is successful in predicting a difference in the degree of bias at the beginning than in the rest of the gene under some circumstances, as observed in *Escherichia coli*, but grossly overestimates the amount of bias expected. G+C composition of the genome is a vital factor for codon usage variation. This variation mostly lies in the third position of the codons (<20% to >90% G+C), as it is immune to changes. GC-rich organisms tend to prefer GC-containing codons over AT-containing ones. Consequently each organism has their optimal and nonoptimal codons.

Bacteria can acquire foreign genes through HGT (horizontal gene transfer). Bacteriophages are the major player in the HGT phenomenon. Bacteriophage can

mobilize genetic material between distantly related bacterial species. At the time of introduction into the recipient host bacterium, the foreign genes tend to retain the characteristics of the donor bacterium and it may vary significantly from the native genes of the recipient bacterium in terms of optimal codon usage. For the detection of the horizontally transferred foreign genes various parameters of the codon usage such as relative adaptiveness (RA), mean difference of RA, codon adaptation index (CAI), synonymous substitution rate (SSR) between codons can be used. CAI is a measure of similarity of a gene's synonymous codon usage to that of a standard set of highly expressed genes for that organism. The mean deference (MD) of the relative adaptiveness (RA) of the codons of the foreign genes from that of the native genes give us a clue about by what extent the foreign gene varies from the native genes in a host bacterium. Here we present a detailed analysis of a selected set of parameters such as RA, MD of RA, CAI, and SSR of the codon usage pattern of the six phage encoded toxin genes. These are cholera toxin, shiga toxin, neurotoxin C1, enterotoxins type A, cytotoxin and diphtheria toxin. To the best of our knowledge these parameters of the codon usage has not been utilized previously in predicting the time of acquisition of foreign genes. In previous method the rate of horizontal gene transfer was estimated, but not their evolutionary time of acquisition. In this study we proposed a hypothesis involving the conceptual relationship between the evolutionary time of acquisition of the foreign genes and the selected set of parameters of the codon usage and adopt a novel approach for the prediction of the comparative time of the acquisition of the foreign genes on the basis of the analyses the selected parameters.

Chapter 3

Mathematical Indices

3.1.1 Codon usage indices

This document describes the indices calculated by CodonW, by default only the G+C content of the sequence is reported. The others being dependent on the genetic code selected. More than one index may be calculated at the same time.

3.1.2 Codon Adaptation Index (CAI) (Sharp and Li 1987).

CAI is a measurement of the relative adaptiveness of the codon usage of a gene towards the codon usage of highly expressed genes. The relative adaptiveness (w) of each codon is the ratio of the usage of each codon, to that of the most abundant codon for the same amino acid. The relative adaptiveness 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. The CAI index is defined as the geometric mean of these relative adaptiveness values. Non-synonymous codons and termination codons (dependent on genetic code) are excluded.

To prevent a codon absent from the reference set but present in other genes from having a relative adaptiveness value of zero, which would cause CAI to evaluate to zero for any genes which used that codon; it was suggested that absent codons should be assigned a frequency of 0.5 when estimating w(Sharp and Li 1987). An alternative suggestion was that w should be adjusted to 0.01 where otherwise it would be less than this value (Bulmer 1988). CodonW does not adjust the w value if a non-zero-input value is found; zero values are assigned a value of 0.01.

3.1.3 Frequency of Optimal codons (Fop) (Ikemura 1981).

This index, is the ratio of optimal codons to synonymous codons (genetic code dependent). 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 Fop index or a modified

Fop index. 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 the modified F op index, negative values are adjusted to zero.

3.1.4 Codon Bias Index (CBI) (Bennetzen and Hall 1982).

Codon bias index is another measure of directional codon bias, it measures the extent to which a gene uses a subset of optimal codons. CBI is similar to Fop as used by Ikemura, with expected usage 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 the number of optimal codons to be less than expected by random change. This results in a negative value for CBI.

3.1.5 The effective number of codons (NC) (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. Initially the homozygosity for each amino acid is estimated from the squared codon frequencies (see Wright 1990).

If amino acids are rare or missing, adjustments must be made. When there are no amino acids in a synonymous family, Nc is not calculated as the gene is either too short or has extremely skewed amino acid usage (Wright 1990). An exception to this is made for genetic codes where isoleucine is the only 3-fold synonymous amino acid, and is not used in the protein gene. The reported value of Nc is always between 20 (when only one codon is effectively used for each amino acid) and 61 (when codons are used randomly). If the calculated Nc is greater than 61 (because codon usage is more evenly distributed than expected), it is adjusted to 61.

3.1.6 G+C content of the gene

□ The frequency of nucleotides that are guanine or cytosine.

3.1.7 G+C content 3rd position of synonymous codons (GC_{3s})

The fraction of codons, that are synonymous at the third codon position, which have

either a guanine of cytosine at that third codon position.

3.1.8 Base composition at silent sites.

Selection of this option calculates four separate indices, i.e. G_{3s}, C_{3s}, A_{3s} & T_{3s}. Although correlated with GC_{3s}, this index is not directly comparable. It quantifies the usage of each base at synonymous third codon positions. When calculating GC_{3s} each synonymous amino acid has at least one synonym with G or C in the third position. Two or three fold synonymous amino acids do not have an equal choice between bases in the synonymous third position. The index A3s is the frequency that codons have an A at their synonymous third position, relative to the amino acids that could have a synonym with A in the synonymous third codon position. The codon usage analysis of *Caenorhabditis elegans* identified a trend correlated with the frequency of G3s. Though it was not clear whether it reflected variation in base composition (or mutational biases) among regions of the *C. elegans* genome, or another factor (Stenico et al. 1994).

3.1.9 Length silent sites (Lsil).

- Frequency of synonymous codons.

3.1.10 Length amino acids (Laa).

- Equivalent to the number of translatable codons.

3.1.11 Hydropathicity of protein

The general average hydropathicity or (GRAVY) score, for the hypothetical translated gene product. It is calculated as the arithmetic mean of the sum of the hydropathic indices of each amino acid (Kyte and Doolittle 1982). This index has been used to quantify the major COA trends in the amino acid usage of *E. coli* genes (Lobry and Gautier 1994).

3.1.12 Aromaticity score

The frequency of aromatic amino acids (Phe, Tyr, Trp) in the hypothetical translated gene product. The hydropathicity and aromaticity protein scores are indices of amino acid usage. The strongest trend in the variation in the amino acid composition

of E. coli genes is correlated with protein hydropathicity, the second trend is correlated with gene expression, while the third is correlated with aromaticity (Lobry and Gautier 1994). The variation in amino acid composition can have applications for the analysis of codon usage. If total codon usage is analysed, a component of the variation will be due to differences in the amino acid composition of genes.

Chapter 4

Methods and materials

4.1 Materials

4.1.1 Bacterial Strains

4.1.1.1 *Vibrio choleare* O1

Vibrio cholerae, a Gram-negative bacterium belonging to the γ-subdivision of the family Proteobacteriaceae is the etiologic agent of cholera, a devastating diarrheal disease which occurs frequently as epidemics. Any bacterial species encountering a broad spectrum of environments during the course of its life cycle is likely to develop complex regulatory systems and stress adaptation mechanisms to best survive in each environment encountered. Toxigenic *V. cholerae*, which has evolved from environmental nonpathogenic V. cholerae by acquisition of virulence genes, represents a paradigm for this process in that this organism naturally exists in an aquatic environment but infects human beings and cause cholera. The *V. cholerae* genome, which is comprised of two independent circular mega-replicons, carries the genetic determinants for the bacterium to survive both in an aquatic environment as well as in the human intestinal environment. Pathogenesis of *V. cholerae* involves coordinated expression of different sets of virulence associated genes, and the synergistic action of their gene products. Although the acquisition of major virulence genes and association between *V. cholerae* and its human host appears to be recent, and reflects a simple pathogenic strategy, the establishment of a productive infection involves the expression of many more genes that are crucial for survival and adaptation of the bacterium in the host, as well as for its onward transmission and epidemic spread. While a few of the virulence gene clusters involved directly with cholera pathogenesis have been characterized, the potential exists for identification of yet new genes which may influence the stress adaptation, pathogenesis, and epidemiological characteristics of *V. cholerae*. Coevolution of bacteria and mobile genetic elements (plasmids, transposons, pathogenicity islands, and phages) can determine environmental survival and pathogenic interactions between bacteria and their hosts. Besides horizontal gene transfer mediated by

genetic elements and phages, the evolution of pathogenic V. cholerae involves a combination of selection mechanisms both in the host and in the environment. The occurrence of periodic epidemics of cholera in endemic areas appear to enhance this process.

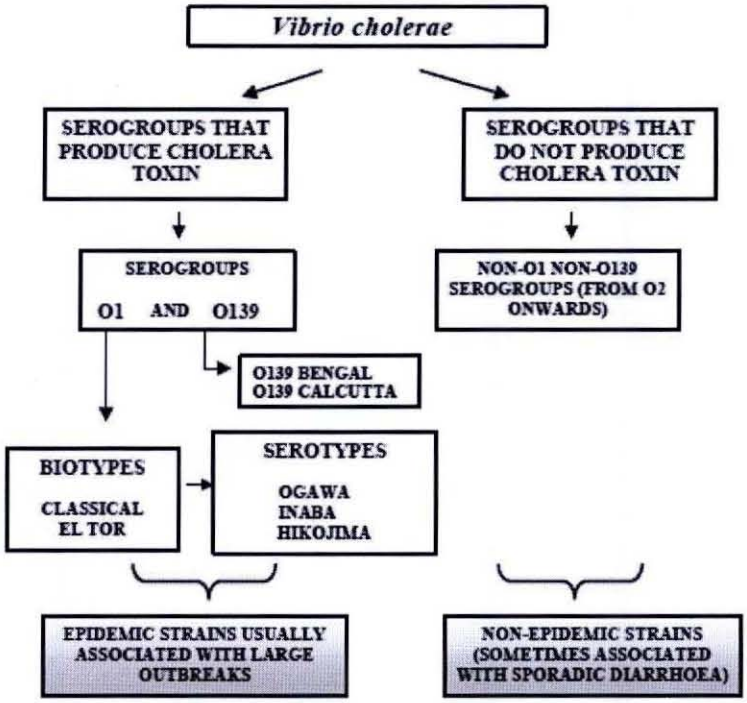


Figure 1: Classification of *Vibrio cholerae* serogroups into epidemic and non-epidemic groups is shown; the respective biotypes within a serotype are also represented. The O1 and O139 serogroups are currently the only ones associated with epidemic cholera.

4.1.1.2 *Pseudomonas aeruginosa*

Pseudomonas aeruginosa is a versatile Gram-negative bacterium that grows in soil, coastal marine habitats, and on plant and animal tissues. Pseudomonas aeruginosa is a ubiquitous environmental bacterium that is one of the top three causes of opportunistic human infections. People with cystic fibrosis, burn victims, and other patients in intensive care units are particularly at risk of disease resulting from P. aeruginosa infection. A major factor in its prominence as a pathogen is its intrinsic resistance to antibiotics and disinfectants. At 6.3 million base pairs, this is the largest bacterial genome sequenced, and the sequence provides insights into the basis of the versatility and intrinsic drug resistance of P. aeruginosa. Consistent with its larger

genome size and environmental adaptability, P. aeruginosa contains the highest proportion of regulatory genes observed for a bacterial genome and a large number of genes involved in the catabolism, transport and efflux of organic compounds as well as four potential chemotaxis systems.

4.1.1.3 *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).

4.1.1.3.1 *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. As expected for a Low G+C species there is a

predominance of A/U ending codons.

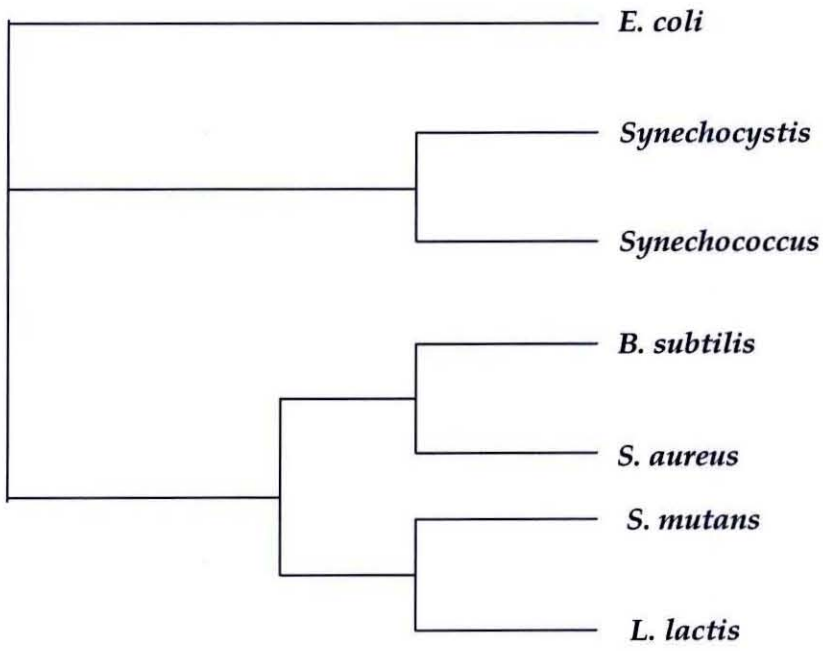


Figure 2: Diagrammatic representation of the relationship between species analysed in this thesis and *E. coli* and *B. subtilis*.

4.1.1.4 *Escherichia coli* O157

Escherichia coli O157:H7 is one of hundreds of strains of the gram-negative bacillus E. coli. Most strains are harmless, colonizing the intestines of healthy humans and animals, where they uppress the growth of pathogenic bacterial species and synthesize appreciable amounts of vitamin K and vitamin B complex. But a few strains cause gastroenteritis in humans by 4 mechanisms: adherence to small-bowel mucosa, direct invasion of mucosal cells, and disruption of the microvillous brush border and toxin release. The class enterohemorrhagic E. coli, which includes E. coli O157:H7, produces hemorrhagic colitis by elaborating one or more cytotoxins closely related to the Shigella toxin. These toxins, variably called Shiga's toxins or verotoxins, damage intestinal epithelium and appear to possess neurotoxic and enterotoxic properties.1 E. coli O157:H7 was not recognized as a human

pathogen until 1982, when the serotype was identified in stool specimens from American patients with bloody diarrhea.2 Since then at least 65 outbreaks of the infection have been reported, 3 most recently in Walkerton, Ont., where at least 7 residents died after drinking contaminated municipal water.4 Most outbreaks occur after people eat undercooked ground beef that is likely contaminated during slaughtering and subsequent meat processing. Outbreaks have also been caused by unpasteurized milk and similar products.

4.1.1.5 *Clostridium botulinum*

Clostridium botulinum (C. botulinum) is a spore-forming bacterium that produces a very powerful neurotoxin that causes botulism. The toxin is among the most toxic of all naturally occurring substances. Botulism is usually associated with consumption of the toxin in food. However, in rare cases the toxin can be produced in infected wounds or in the intestinal tracts of young infants.

C. botulinum spores can be found in soil and are very resistant to heat and other treatments. Because naturally occurring levels of spores are low, growth is required to produce toxin. *C. botulinum* grows under anaerobic (no oxygen) conditions.

4.1.1.6 *Corynebacterium diphtheriae*

Corynebacteria are Gram-positive, aerobic, nonmotile, rod-shaped bacteria classified as Actinobacteria. Corynebacteria are related phylogenetically to mycobacteria and actinomycetes. They do not form spores or branch as do the actinomycetes, but they have the characteristic of forming irregular, club-shaped or V-shaped arrangements in normal growth. They undergo snapping movements just after cell division, which brings them into characteristic forms resembling Chinese letters or palisades.

The genus Corynebacterium consists of a diverse group of bacteria including animal and plant pathogens, as well as saprophytes. Some corynebacteria are part of the normal flora of humans, finding a suitable niche in virtually every anatomic site, especially the skin and nares. The best known and most widely studied species is Corynebacterium diphtheriae, the causal agent of the disease diphtheria.

Diphtheria is an upper respiratory tract illness characterized by sore throat, low fever, and an adherent membrane (called a pseudomembrane on the tonsils, pharynx, and/or nasal cavity. Diphtheria toxin produced by C. diphtheriae, can cause myocarditis, polyneuritis, and other systemic toxic effects. A milder form of diphtheria can be restricted to the skin.

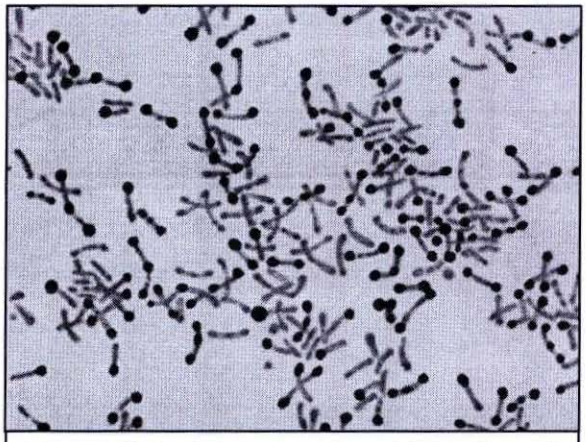


Figure 3: Stained *Corynebacterium* cells. The "barred" appearance is due to the presence of polyphosphate inclusions called metachromatic granules. Note also the characteristic "Chinese-letter" arrangement of cells.

Diphtheria is a contagious disease spread by direct physical contact or breathing aerosolized secretions of infected individuals. Once quite common, diphtheria has largely been eradicated in developed nations through wide-spread use of the DPT vaccine. For example, in the U.S., between 1980 and 2004 there were 57 reported cases of diphtheria. However, it remains somewhat of a problem worldwide (3,978 reported cases to WHO in 2006) in the face of efforts to achieve global vaccination coverage.

Diphtheria is a serious disease, with fatality rates between 5% and 10%. In children under 5 years and adults over 40 years, the fatality rate may be as much as 20%. Outbreaks, although very rare, still occur worldwide, even in

developed nations. Following the breakup of the former Soviet Union in the late 1980s, vaccination rates in the constituent countries fell so low that there was a surge in diphtheria cases. In 1991 there were 2,000 cases of diphtheria in the USSR. By 1998, according to Red Cross estimates, there were as many as 200,000 cases in the Commonwealth of Independent States, with 5,000 deaths.

4.1.2 Toxin Genes

4.1.2.1 Cholera toxin

Cholera toxin (sometimes abbreviated to CTX, Ctx, or CT) is a protein complex secreted by the bacterium *Vibrio cholerae*. CTX is responsible for the harmful effects of cholera infection.

4.2.2.1.1 Cholera toxin Structure

The cholera toxin is an oligomeric complex made up of six protein subunits: a single copy of the A subunit (part A), and five copies of the B subunit (part B). The two parts are connected by a disulfide bond. The three-dimensional structure of the toxin was determined using X-ray crystallography by Zhang et al. in 1995.

The five B subunits—each weighing 12 kDa, and all coloured blue in the accompanying figure—form a five-membered ring. The A subunit has two important segments. The A1 portion of the chain (CTA1, red) is a globular enzyme payload that ADP-ribosylates G proteins, while the A2 chain (CTA2, orange) forms an extended alpha helix which seats snugly in the central pore of the B subunit ring.

This structure is similar in shape, mechanism, and sequence to the heat-labile enterotoxin secreted by some strains of the Escherichia coli bacterium.

4.1.2.1.2 Mechanism of poisonous action on humans

The pentameric part B of the toxin molecule binds to the surface of the intestinal epithelium cells. Part A detaches from the pentameric part upon binding, and gets inside the cell via receptor-mediated endocytosis. Once inside the cell, it permanently ribosylates the Gs alpha subunit of the heterotrimeric G protein resulting in constitutive cAMP production. This in

turn leads to secretion of H2O, Na+, K+, Cl-, and HCO3- into the lumen of the small intestine resulting in rapid dehydration.

4.1.2.1.3 Origin of Cholera toxin

The gene encoding the cholera toxin is introduced into V. cholerae by horizontal gene transfer. Virulent strains of *V. cholerae* carry a variant of lysogenic bacteriophage called CTXf or CTXφ.

4.1.2.1.4 Working Mechanism of Cholera toxin

Once secreted, the B subunit ring of CTX will bind to GM1 gangliosides on the surface of the host's cells. After binding takes place, the entire CTX complex is internalised by the cell and the CTA1 chain is released by the reduction of a disulfide bridge.

CTA1 is then free to bind with a human partner protein called ADP-ribosylation factor 6 (Arf6); binding to Arf6 drives a change in the conformation (the shape) of CTA1 which exposes its active site and enables its catalytic activity.

The CTA1 fragment catalyses ADP ribosylation from NAD to the regulatory component of adenylate cyclase, thereby activating it. Increased adenylate cyclase activity increases cyclic AMP (cAMP) synthesis causing massive fluid and electrolyte efflux, resulting in diarrhea.

4.1.2.1.5 The Actions of Cholera Toxin

When cholera toxin is released from the bacteria in the infected intestine, it binds to the intestinal cells known as enterocytes (epithelial cell in above diagram) through the interaction of the pentameric B subunit of the toxin with the GM1 ganglioside receptor on the intestinal cell, triggering endocytosis of the toxin. Next, the A/B cholera toxin must undergo cleavage of the A1 domain from the A2 domain in order for A1 to become an active enzyme. Once inside the enterocyte, the enzymatic A1 fragment of the toxin A subunit

enters the cytosol, where it activates the G protein Gsa through an ADP-ribosylation reaction that acts to lock the G protein in its GTP-bound form, thereby continually stimulating adenylate cyclase to produce cAMP. The high cAMP levels activate the cystic fibrosis transmembrane conductance regulator (CFTR), causing a dramatic efflux of ions and water from infected enterocytes, leading to watery diarrhoea.

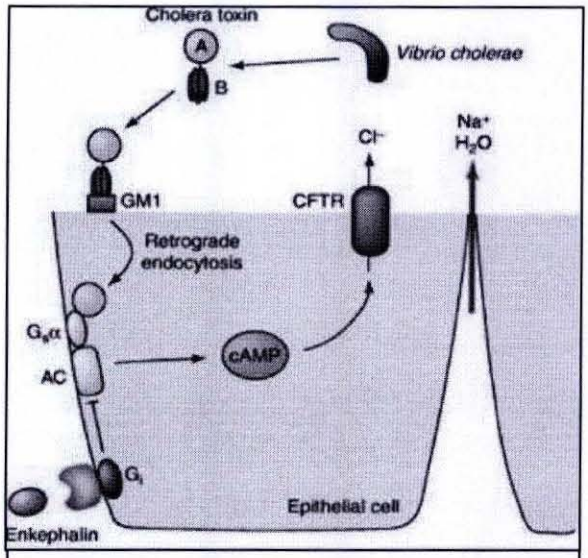


Figure 4: A, B (cholera toxin subunits); GM1 (GM1 ganglioside receptor); G_sa (G protein); AC (adenylate cyclase); G_i (G protein); cAMP (cyclic AMP); CFTR (cystic fibrosis transmembrane conductance regulator)

One area of anti-diarrhoea treatment lies in the stimulation of enkephalins, which regulate intestinal secretion by acting directly on enterocytes. Enkephalins bind to the opioid receptors on enterocytes, which act through G proteins to inhibit the stimulation of cAMP synthesis induced by cholera toxin, thereby directly controlling ion transport.

4.1.2.1.6 Applications of Cholera toxin

Because the B subunit appears to be relatively non-toxic, researchers have found a number of applications for it in cell and molecular biology. It is routinely used as a neuronal tracer.

GM1 gangliosides are found in lipid rafts on the cell surface. B subunit complexes labelled with fluorescent tags or subsequently targeted with antibodies can be used to identify rafts.

4.1.2.1.7 Diversity in Cholera Strains

The effects of cholera involve the actions of other *Vibrio cholerae* toxins that aid the pathogen in its colonisation, coordinated expression of virulence factors,

and toxin action. These additional proteins include zona occludens toxin (zot, involved in Vibrio cholerae invasion by acting to decrease intestinal tissue resistance), accessory cholera toxin (ace, increases fluid secretion), toxin-coregulated pilus (tcpA, essential colonisation factor and receptor for the CTXf phage), NAG-specific heat-labile toxin (st), and outer membrane porin proteins (ompU and ompT). The expression of virulence factors is controlled by the transcriptional factors ToxR, TcpP and ToxT. Different strains of *Vibrio cholerae* produce differing sets and amounts of these auxiliary toxins, which in turn affect the clinical symptoms of cholera and its responsiveness to treatment.

For example, the cholera outbreak in Russia in 1942 was caused by the El Tor biotype strain of *Vibrio cholerae*, rather than the classical biotype that caused the pandemics in the 19th and early 20th centuries. The El Tor biotype can carry several extra copies of CTXf bacteriophage that contains the toxin genes ctxAB (encodes cholera toxin A and B subunits), zot (encodes zona occludens toxin) and ace (encodes accessory cholera toxin), leading to an increase in cholera toxin production. The El Tor biotype can also produce haemolysin, which is capable of lysing red blood cells by attacking their membranes. In addition, unlike the classical biotype, the El Tor biotype generates novel toxin strains through CTXf phage conversion. These El Tor strains produce different, milder clinical symptoms, with many patients showing asymptomatic cholera not accompanied by dehydration.

4.1.2.2 Shiga toxin

Shiga toxins are a family of related toxins with two major groups, Stx1 and Stx2, whose genes are considered to be part of the genome of lambdoid prophages. The toxins are named for Kiyoshi Shiga, who first described the bacterial origin of dysentery caused by Shigella dysenteriae. The most common sources for Shiga toxin are the bacteria S. dysenteriae and the

Shigatoxigenic group of Escherichia coli (STEC), which includes serotype O157:H7 and other enterohemorrhagic E. coli.

4.1.2.2.1 Nomenclature of Shiga toxin

There are many terms that microbiologists use to describe Shiga toxin and differentiate between different forms of it. Many of these terms are used interchangeably.

1. Shiga toxin (Stx) - true Shiga toxin is produced by Shigella dysenteriae.
2. Shiga-like toxin 1 and 2 (SLT-1 and 2 or Stx-1 and 2) - the Shiga toxins produced by some E. coli strains. Stx-1 differs from Stx by only 1 amino acid. Stx-2 shares 56% sequence homology with Stx-1.
3. Cytotoxins - an archaic denotation for Stx, used in a broad sense.
4. Verocytotoxins - a seldom used denotation for Stx, from the hypersensitivity of Vero cells to Stx.

4.1.2.2.2 Structure of Shiga toxin

The toxin has two subunits—designated A and B—and is one of the AB_5 toxins. The B subunit is a pentamer that binds to specific glycolipids on the host cell, specifically globotriaosylceramide (Gb_3). Following this, the A subunit is internalised and cleaved into two parts. The A1 component then binds to the ribosome, disrupting protein synthesis. Stx-2 has been found to be approximately 400 times more toxic (as quantified by LD50 in mice) than Stx-1.

Gb_3 is, for unknown reasons, present in greater amounts in renal epithelial tissues, to which the renal toxicity of Shiga toxin may be attributed. Gb3 is also found in CNS neurons and endothelium, which may lead to neurotoxicity.

The toxin requires highly specific receptors on the cells' surface in order to attach and enter the cell; species such as cattle, swine, and deer which do not

carry these receptors may harbor toxigenic bacteria without any ill effect, shedding them in their feces, from where they may be spread to humans.

4.1.2.2.3 Mechanism

Shiga toxins act to inhibit protein synthesis within target cells by a mechanism similar to that of ricin toxin produced by Ricinus communis. After entering a cell, the protein functions as an N-glycosidase, cleaving several nucleobases from the RNA that comprises the ribosome, thereby halting protein synthesis.

4.1.2.3 Neurotoxin C1

Botulinum neurotoxins (BoNTs) are produced by Clostridium botulinum and cause the neuroparalytic syndrome of botulism. With a lethal dose of 1 ng/kg, they pose a biological hazard to humans and a serious potential bio-weapon threat. On the other hand, BoNTs have become a powerful therapeutic tool in the treatment of a variety of neurological, ophthalmic, and other disorders manifested by abnormal, excessive, or inappropriate muscle contractions. Experimental studies are also underway that explore the use of BoNTs in the management of chronic pain, such as headache and migraine. BoNTs bind with high specificity at neuromuscular junctions and they impair exocytosis of synaptic vesicles containing acetylcholine through specific proteolysis of SNAREs which constitute part of the synaptic vesicle fusion machinery. The molecular details of the toxin-cell recognition have been elusive.

4.1.2.3.1 Biochemical mechanism of toxicity

The heavy chain of the toxin is particularly important for targeting the toxin to specific types of axon terminals. The toxin must get inside the axon terminals in order to cause paralysis. Following the attachment of the toxin heavy chain to proteins on the surface of axon terminals, the toxin can be taken into neurons by endocytosis. The light chain is able to cleave

endocytotic vesicles and reach the cytoplasm. The light chain of the toxin has protease activity. The type A toxin proteolytically degrades the SNAP-25 protein, a type of SNARE protein. The SNAP-25 protein is required for the release of neurotransmitters from the axon endings. Botulinum toxin specifically cleaves these SNAREs, and so prevents neuro-secretory vesicles from docking/fusing with the nerve synapse plasma membrane and releasing their neurotransmitters.

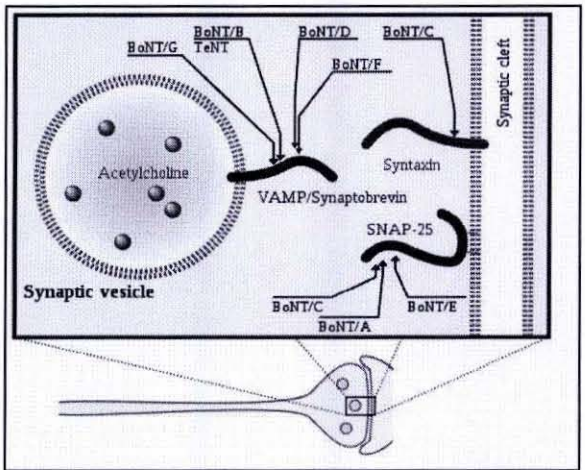


Figure 5: Target molecules of botulinum (BoNT) and tetanus (TeNT) toxins inside the axon terminal.

Though it affects the nervous system, common nerve agent treatments (namely the injection of atropine and 2-pam-chloride) will increase mortality by enhancing botulin toxin's mechanism of toxicity. Attacks involving botulinum toxin are distinguishable from those involving nerve agent in that NBC detection equipment (such as M-8 paper or the ICAM) will not indicate a "positive" when a sample of the agent is tested. Furthermore, botulism symptoms develop relatively slowly, over several days compared to nerve agent effects, which can be instantaneous.

4.1.2.4 Enterotoxins type A

Superantigens (SAgs) are bacterial and viral proteins that share the ability to activate a large fraction of T-lymphocytes (Marrack and Kappler, 1990). The staphylococcal enterotoxins are the best characterized of the SAgs. They have been shown to bind as unprocessed proteins to major histocompatibility complex (MHC) class II molecules on antigen presenting cells, and

subsequently activate T-cells through interaction with the variable region of the T-cell receptor a-chain (TCR-Vp) encoded by certain families of TCR-Vf genes (Marrack and Kappler, 1990). This results in the activation of between 2 and 15% of all T-cells ultimately leading to proliferation, production of a variety of cytokines as well as expression of cytotoxic activity (see review edited by Mbller, 1993).

Staphylococcal enterotoxins (SEs) are a major cause of food poisoning and bacterial Gram-positive shock in humans. Excessive induction of cytokines has been implicated as a central pathogenic factor in SAg-related toxicity.

SEs can be divided into two groups based on sequence homology comprising SEA/SED/SEE and SEB/SEC 1-3 (Marrack and Kappler, 1990; Ren et al., 1994). The sequence homology ranges from 25 to 83% with SEA and SEE being the most closely related (Betley et al., 1992). In addition, SEA and SEE have been shown to share the ability to bind zinc which has been proposed to be crucial for their interaction with MHC class II (Fraser et al., 1992).

Recent studies have suggested that SEA has two distinct MHC class II binding regions (Hedlund et al., 1991; Betley et al., 1992; Fraser et al., 1992; Abrahmsen et al., 1995). These consist of a moderate affinity site, which is Zn2+ dependent, and a lower affinity site which resembles the binding site found in SEB. Alanine substitution mutagenesis of SEA has revealed that the two binding sites cooperate to form a strong MHC class 1I-SEA interaction (Abrahmsen et al., 1995).

4.1.2.5 Cytotoxin

Pseudomonas aeruginosa is an opportunistic pathogen that is capable of producing life-threatening disease in immunocompromised individuals. Those who are especially at risk include patients with severe burns, cancer, diabetes, or cystic fibrosis. In those with cystic fibrosis, *P. aeruginosa* can cause

persistent lung infections, indicating that the host immune system is incapable of clearing the bacteria. As macrophages represent one of the primary lines of defense against infections, it has been suggested that these phagocytes may not be functioning correctly in the lungs of patients with cystic fibrosis. One of the ways in which *P. aeruginosa* may protect itself from such basic host defenses is through production of a cytotoxin. *P. aeruginosa* cytotoxin, previously named leukocidin, has been isolated from autolysates of *P. aeruginosa* cells and appears to be associated with all isolates of *P. aeruginosa*. It inactivates eucaryotic cells by forming lesions or pores in the membrane of target cells of the immune system. This results in increased plasma membrane permeability to small molecules and ions. Such intoxication has been documented in granulocytes, endothelial cells, Ehrlich ascites tumor cells, and human leukemic cells. In the case of granulocytes, treatment with the cytotoxin causes an inhibition of the ability of the granulocytes to kill *P. aeruginosa* cells. The present study was designed to determine the bacterial cellular localization of cytotoxin and to examine its effect on macrophages. Toward this end, various bacterial cell compartments were tested for the presence of cytotoxin, and osmotic shock fluid (periplasmic contents) and a purified preparation of cytotoxin were observed for their interaction with mouse macrophage cell line P388D1. Results of previous studies have indicated that this cell line is an appropriate model for unelicited mouse peritoneal macrophages and cultured human peripheral blood monocytes in the assessment of opsonized phagocytosis of *P.aeruginosa*.

4.1.2.6 Diphtheria toxin

Diphtheria toxin is an exotoxin secreted by Corynebacterium diphtheriae, the pathogen bacterium that causes diphtheria.

4.1.2.6.1 Structure

Diphtheria toxin is a single polypeptide chain of 535 amino acids consisting of two subunits linked by disulfide bridges. Binding to the cell surface of the less

stable of these two subunits allows the more stable part of the protein to penetrate the host cell.

4.1.2.6.2 Mechanism

It catalyzes the ADP-ribosylation of eukaryotic elongation factor-2 (eEF2), inactivating this protein. It does so by ADP-ribosylating the unusual amino acid diphthamide. In this way, it acts as a RNA translational inhibitor.

The exotoxin A of Pseudomonas aeruginosa uses a similar mechanism of action.

4.1.3. Softwares utilized for analysis

4.1.3.1 The Institute for Genome Research (TIGR)

The Institute for Genomic Research (TIGR) was a non-profit genomics research institute founded in 1992 by Craig Venter in Rockville, Maryland, United States. It is now a part of the J. Craig Venter Institute.

TIGR sequenced the first genome of a free-living organism, the bacterium Haemophilus influenzae, in 1995. This landmark project, led by TIGR scientist Robert Fleischmann, led to an explosion of genome sequencing projects, all using the whole-genome sequencing technique pioneered earlier but never used for a whole bacterium until TIGR's project. TIGR scientist Claire Fraser led the projects to sequence the second bacterium, Mycoplasma genitalium in 1996, and less than a year later TIGR's Carol Bult led the project to sequence the first genome of an Archaeal species, Methanococcus jannaschii. TIGR followed these accomplishments with the genomes of the pathogenic bacteria Borrelia burgdorferi (which causes Lyme Disease) in 1997, and Treponema pallidum (which causes syphilis) in 1998. In 1999 TIGR published the sequence of the radioresistant polyextremophile Deinococcus radiodurans.

TIGR went on to become the world's leading center for microbial genome sequencing, and it also participated in the Human Genome Project and many other genome projects. Its bioinformatics group developed many of the pioneering software algorithms that were used to analyze these genomes, including the automatic gene finder GLIMMER and the genome alignment program MUMmer.

Following the 2001 anthrax attacks, TIGR partnered with the National Science Foundation and the FBI to sequence the strain of Bacillus anthracis used in those attacks. The results of this analysis were published in the journal Science in 2002[2]. The genetic evidence was later credited by the FBI with

helping to pinpoint the precise sample of anthrax bacteria, from a lab in Fort Detrick, Maryland, that was the source of the attacks.

TIGR's Genome Projects are a collection of curated databases containing DNA and protein sequence, gene expression, cellular role, protein family, and taxonomic data for microbes, plants and humans. The access to the data is facilitated by TIGR's Internet2 high-speed research network connection which is supported in part by the National Science Foundation under grant ANI-0333537. Anonymous FTP access to sequence data is also provided.

4.1.3.2 Graphical codon usage analyzer (GCUA)

Differences in codon usage preference among organisms lead to a variety of problems concerning heterologous gene expression but can be overcome by rational gene design and gene synthesis. The gcua tool displays the codon quality either in codon usage frequency values or relative adaptiveness values.

4.1.3.3 EMBOSS (European biology open software suite)

EMBOSS is "The European Molecular Biology Open Software Suite". EMBOSS is a free Open Source software analysis package specially developed for the needs of the molecular biology (e.g. EMBnet) user community. The software automatically copes with data in a variety of formats and even allows transparent retrieval of sequence data from the web. Also, as extensive libraries are provided with the package, it is a platform to allow other scientists to develop and release software in true open source spirit. EMBOSS also integrates a range of currently available packages and tools for sequence analysis into a seamless whole. EMBOSS breaks the historical trend towards commercial software packages.

The uses and interfaces to EMBOSS have long grown beyond our ability to keep track of them. EMBOSS is used extensively in production environments

rather than being the sort of "research project" code that gets presented at conferences, but never actually deployed.

EMBOSS has several important advantages:

- A properly constructed toolkit for creating robust bioinformatics applications or workflows.
- A comprehensive set of sequence analysis programs.
- All sequence and many alignment and structural formats are handled.
- Extensive programming library for common sequence analysis tasks.
- Additional programming libraries for many other areas including string handling, pattern-matching, list processing and database indexing.
- It is free-of-charge.
- It is an open-source project.
- It runs on practically every UNIX you can think of and some that you can't, plus MS Windows and MacOS.
- Each application has the same style of interface so master one and you've mastered them all.
- The consistent user interface facillitates GUI designers and developers.
- It integrates other popular publicly available packages.
- It is free of arbitrary size limits: there are no limits on the amount of data that can be processed. For the programmer, memory management for objects such as sequences and arrays is simplified

4.1.3.4 JCat (Java Codon Adaptation Tool)

The CodonAdaptationTool (JCAT) presents a simple method to adapt the Codon Usage to most sequenced prokaryotic organisms and selected eukaryotic organisms. The codon adaptation plays a major role in cases where foreign genes are expressed in hosts and the codon usage of the host differs

from that of the organism where the gene stems from. Unadapted codons in the host can for example lead to a minor expression rate.

The adaptation is based on CAI-values proposed by PM Sharp et.al. The CAI-values were calculated by applying an algorithm from A Carbone et. al. The eukaryotic genomes of mouse and human contain different kinds of biases along the chromosomes and the algorithm is not perfectly suited for this problem. Results in this field should be handled with care. The mean codon usage for a certain organism was derrived by summing over all CAI-values of all genes of this organism (except genes without an amino acid sequence, e.g. RNAs) devided by the number of genes. This data is also presented in the graphical output of the codon adaptation.

As a further option for the codon adaptation the opportunity to avoid rho-independent trancription terminators is provided. The algorithm for the prediction of these structures is based on a model from MD Ermolaeva et. al. Another feature is the possibility to avoid restriction enzyme binding sites in the adapted DNA. The data for the restriction enzymes was therefor derrived from the "The Restriction Enzyme Database" (REBASE).

4.2 Methods

4.2.1. Constructing the codon usage tables for toxin genes and corresponding host bacteria:

The nucleotide sequence of the six toxin genes were retrieved from the TIGR (The institute of genome research) database (www.tigr.org). Codon usage table with the relative codon frequencies of individual codons for each of the toxin gene was constructed by the The Sequence Manipulation Suite web tool (14). The codon usage tables of selected set of host bacteria (*Vibrio choleare* O1, *P. aeruginosa, S. aureus, E.coli* O157, *C. botulinum, C. diphtheriae*) which harbor the toxin genes to be analyzed, were retrieved from the codon usage databse of CUTG (Codon Usage Tabulated from GenBank (ftp distribution) : Codon usage tables for NCBI listed organisms.

4.2.2 Graphical analyses of relative adaptiveness of codon usage frequencies:

Graphical codon usage analyzer (GCUA) web tool was utilized for comparing the codon usage tables of a particular toxin gene and the host bacterium in which the toxin gene resides. Here the bar diagrams were generated based on the relative adaptiveness of codon frequencies. The basic principle for deriving relative adaptiveness values out of codon usage frequency values is the following: for each amino acid the codon with the highest frequency value is set to 100% relative adaptiveness. All other codons for the same amino acid are scaled accordingly. The graphical codon usage analyzer tool is accessible under **http://gcua.schoedl.de/**.

4.2.3 GC content analysis of toxin genes and corresponding host bacterial genome:

The GC content of the toxin gene nucleotide sequences and the bacterial genome were established by the EMBOSS (European biology open software suite). Third letter G+C % was also counted by the web tools of the same package.

4.2.4 Calculating Codon adaptation index:

The CAI value indicates the expressivity of a given gene. It is also useful to identify the poorly expressed genes. For the calculation of CAI each codon is given a weight with respect to the subset of highly expressed genes defined for the considered organism. JCat, (Java Codon Adaptation Tool), a very rapid and easy method for the estimation of CAI, was employed in this study to calculate the CAI values of the six toxin genes.

Appendix 1

$Wi = fi / \max \{fj, \text{all synonymous for i}\}$,

$CAI\ (gene) = L \sqrt[L]{\prod_{i=1}} Wi$

$$\ln CAI\ (gene) = \sum_{i=1}^{64} gi \ln Wi = (\ln Wi)\ gene$$

fi= Frequency of Codon I, calculated over reference set S

L= Number of all codons in a gene

gi = Frequency of codon I on a gene

4.2.5 Estimating sysnonymous substitution rate in the toxin genes:

Substitution rate at the different synonymous sites in the toxin genes were calculated by a computational method developed Adam Eyre-Walker and Michael Bulmer et al. This method was slightly modified to calculate the SSR in the phage encoded toxin genes with respect to their corresponding host genome.

Appendix 2

Let us consider Ci and Cj are relative frequencies of two synonymous codons for a particular amino acid. For toxin genes Ci and Cj is denoted by Citox and Cjtox where as in the host bacteria these are called Cibac and Cjbac Substitution rate between these two codons by the following formula,

$S = -b \ln (1 - p / b)$

Where,

$b = 1 - \{(f1 + f2) (f1 + f3) / n2 - (f3 + f4)(f2 + f4) / n2\}$ [f1= Citox / Cibac , Citox / Cjbac , Cjtox / Cibac , Cjtox / Cjbac]; p= (f3 + f 3) / n where n= Sum of all the four frequencies.

5.1 Comparison of Relative adaptiveness of toxin gene codons:

Relative adptiveness (RA) of the codon frequencies of the six phage encoded toxin genes with respect to their corresponding host genome were determined. The comparison was drawn between the RAs of both toxin gene codons and overall codon frequencies of the host bacteria and mean variance of the RA for each toxin gene from their host was calculated. The difference between the RAs of each codons of the toxin genes and corresponding hosts were shown in figure 6.

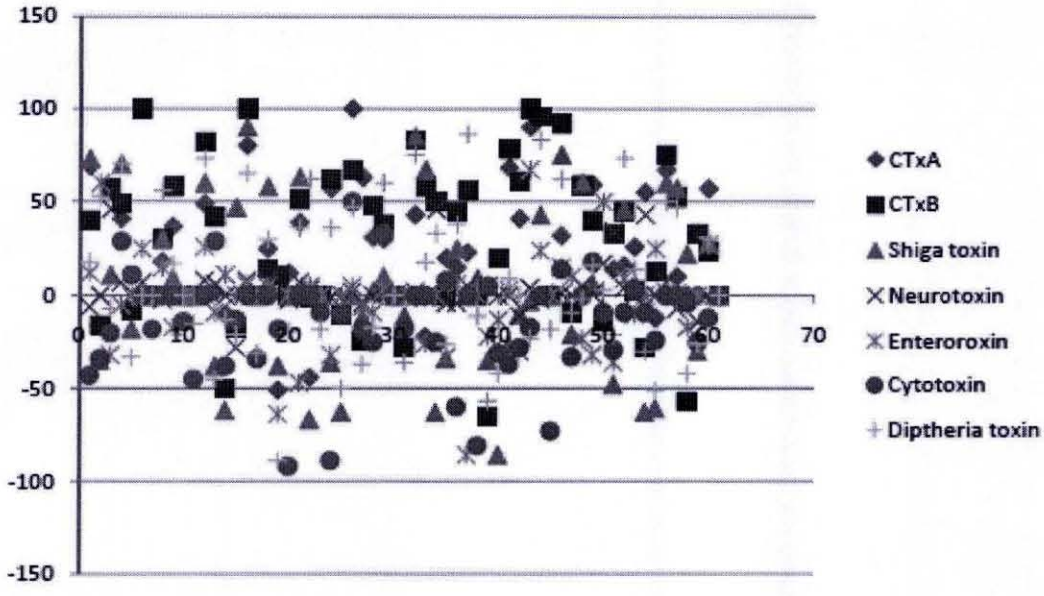


Figure 6: Difference in RA of the each codons between toxin gene and host bacteria: The difference of the relative adaptiveness (RA) of the codons between toxins genes and respective bacteria was calculated and is shown scatter plot. When RA of Codon toxin gene < RA of the Codon host bacteria, the difference was considered as positive values where as negative values were obtained when Codon toxin gene > RA of the Codon host bacteria. Each symbol in the graph represents RA difference for a particular toxin. Each point in the scatter plot represents the difference in RA for a particular codon and the corresponding host bacterium.

Form figure 6 it is evident that the differences of the RAs are greater for shigatoxin, diphtheria toxin and cytotoxin from their respective hosts where as difference in the RA is low for neurotoxin and cytotoxin. Mean difference (MD) between the RA of toxin genes and respective host bacterium was also calculated. Lowest MD was found for neurotoxin which implies its similarities of with the host bacterium *Clostridium botulinum*. High MD was seen for shiga toxin, Diphtheria toxin and cholera toxin. Among these, MD for diphtheria toxin was highest. (Table 1).

Table *1:*

Gene	Total no. of codons	Host organism	CAI	Mean difference of RA	GC content in %	GC content of host	No of codons with GC at 3rd position	3rd letter GC% in the toxin gene	3rd letter GC% in the total CDs in the host
Cholera toxin	384	*Vibrio cholerae*	0.31	31.5	35.27	47.488	90	27.81	44.68
Shiga toxin	316	*E.coli* O157	0.16	35.03	41.56	50.40	103	31	55.17
Neurotoxin C1	1292	*C. botulinum*	0.63	5.91	26.16	28.21	1292	12.77	14.21
Cytotoxin	287	*P. aeruginosa*	0.25	22.53	53.77	66.56	198	68	87.15
Enterotoxin Type A	258	*S. aureus*	0.38	15.61	31.13	32.85	61	22.68	23.64
Diptheria Toxin	561	*C. diphtheriae*	0.15	35.73	42.54	53.48	192	34.22	56.12

Table 1: CAI values and GC content and of the toxin genes

5.2 Analysis and comparison of GC content of toxin genes with corresponding host organisms:

GC content of the six phage encoded toxin genes was estimated and compared with the total GC content of the specific set of bacteria in which they occur. Significant similarity was observed for two toxins (neurotoxin and enterotoxin) with their hosts in terms of G+C percentage. The G+C% of four toxin genes, (Shiga, Cytotoxin, cholera toxin and diphtheria toxin)

significantly lower than their corresponding host bacterium (Table 1). The 3rd letter GC% in the toxin gene was also compared with that of the corresponding host bacteria (Table 1). For neurotoxin and enterotoxin the 3rd letter GC% are very similar to their host bacteria where the rest of the toxin genes showed variation in the 3rd letter GC% when compared with their respective bacteria (Table 2).

Table 2.

Gene	Total no. of codons	Host organism	CAI	Mean difference of RA	GC content in %	GC content of host	No of codons with GC at 3rd position	3rd letter GC% in the toxin gene	3rd letter GC% in the total CDs in the host
Cholera toxin	384	*Vibrio cholerae*	0.31	31.5	35.27	47.488	90	27.81	44.68
Shiga toxin	316	*E.coli* O157	0.16	35.03	41.56	50.40	103	31	55.17
Neurotoxin C1	1292	*C. botulinum*	0.63	5.91	26.16	28.21	1292	12.77	14.21
Cytotoxin	287	*P. aeruginosa*	0.25	22.53	53.77	66.56	198	68	87.15
Enterotoxin Type A	258	*S. aureus*	0.38	15.61	31.13	32.85	61	22.68	23.64
Diptheria Toxin	561	*C. diphtheriae*	0.15	35.73	42.54	53.48	192	34.22	56.12

Table 2: CAI values and GC content and of the toxin genes

5.3 Estimating Codon Adaptation Index (CAI):

The codon adaptation index was calculated by Jcat web tool for each of six toxin genes and shown in table 1. On the basis of CAI values genes can be categorized into 4 classes: very highly expressed genes (CAI>0.6), highly expressed genes (0.5>CAI>0.6), moderately expressed genes and (0.35>CAI>0.5) and weakly expressed genes (CAI<0.35). Neurotoxin C1 (CAI: 0.63) and enterotoxin type A (CAI: 0.38) genes can be considered as very highly expressed and moderately expressed genes respectively. CAI value of the rest of the toxin genes were below 0.35 and can be considered as weakly expressed genes. Diphtheria and shiga toxins have the two lowest CAI values.

5.4 Calculating synonymous substitution rate (SSR)

SSR between eight pairs of synonymous codons (encoding eight amino acids: Ala, Leu, Arg, Gln, Gly, Lys, Phe and Ile) were calculated (Table 2). These amino acids were selected for the SSR calculation, because significant difference in the RA of the codons was observed for these amino acids. Among the eight substitutions, two were transversion and the rest were transition type. The geometric mean of the SSR for one particular toxin was determined (Table 3). The highest SSR was seen for neurotoxin C1. Enterotoxin and cytotoxin gene relatively value of SSR where as shiga toxin, diphtheria toxin and cholera toxin exhibited similar values.

Table 3.

substitution	Substitution mode	Corresponding Amino acid	**Substitution Rate (SR)**					
			CTxA in the genetic background of *Vibrio cholerae*	Shiga toxin in the genetic Background of *E.coli* O157	Cytotoxin in the genetic background of *P.aeruginosa*	Enterotoxin in the genetic background of *S. aureus*	Diptheria toxin in the genetic background of *C.diphtheriae*	Neurotoxin in the genetic background of *C.botulinum*
GCT→GCC	Transition	Ala	0.79	0.77	1.61	1.26	0.86	1.96
TTA →TTG	Transition	Leu	0.79	0.73	2.39	0.80	0.79	1.47
CGG→CGT	Tranversion	Arg	*	0.74	1.29	*	0.89	1.08
CAA→CAG	Transition	Gln	0.93	0.82	0.94	0.62	0.96	1.52
GGT→GGC	Transition	Gly	0.95	0.85	1.42	1.17	0.78	2.37
ATA→ATC	Tranversion	Ile	0.70	0.65	2.57	0.71	0.89	2.45
AAA→AG	Transition	Lys	0.65	0.63	0.60	1.27	0.78	1.37
TTT→TTC	Transition	Phe	0.73	1.18	1.69	1.44	0.83	2.35
Mean SR			0.7968	0.8024	1.56	1.039	0.817	1.82

*indicates nul SSR values, Cholera toxin and Enterotoxin do not have CGG codon and hence SSR values could not be calculated.

Table 3: Synonymous substitution rate (SSR) between the eight pairs of synonymous codons in the toxin genes.

5.5 Predicting the evolutionary time of acquisition of phage encoded toxin gene.

We hypothesized that the CAI, MD and SSR of the codon of the toxin genes should correlate with their evolutionary time of acquisition. We proposed that CAI and SSR should have a linear relationship with the time of acquisition of

the toxin gene, whereas the MD of the relative adaptiveness is inversely correlated with the evolutionary acquisition time (Fig 2). In line of our proposed hypothesis we estimated the relative time of acquisition of the six phage encoded toxin genes by their respective host bacteria. By considering all the parameters, we concluded that neurotoxin C1 and Enterotoxin type A was acquired in the distant past by *Clostridium botulinum* and *Staphylococcus aureus*. On the other hand shiga toxin and diphtheria toxin were introduced in the genome of *E.coli* and *C.diptheriae* respectively in relatively recent past. Cholera toxin and cytotoxin were acquired in between the previous events of acquisition.

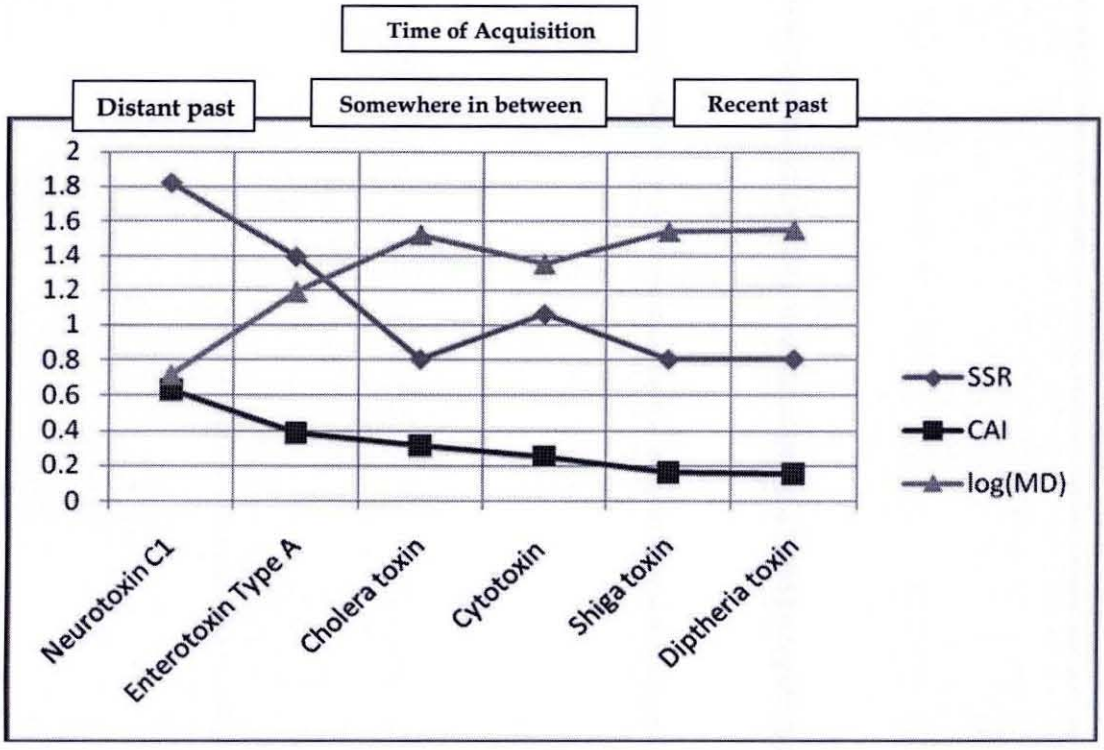


Figure 7(Chakraborty et al): Comparison of the evolutionary time of acquisition of toxin genes: The span of evolutionary time, in which the six toxin genes were introduced into their corresponding host bacteria, is divided into 3 zones: Distant past, Recent past, and time span between distant and recent past. This graph was based on three parameters (CAI, SSR and log (MD) of the toxin genes. The circles represent the different zones of evolutionary time.

Chapter 6

Discussion

Detail analyses of the CAI and RA, MD of the selected set of six phage encoded toxin genes showed that neurotoxin C1 and enterotoxins type A fits well in the translational system in their respective host as they show high CAI values and low MD. On the other hand shiga toxin and diphtheria toxin have low CAI but high MD values. This implies that these two toxins are weakly expressed as they show major difference in codon usage with host bateria. The remaining two toxins showed relatively medium CAI and MD value. The synonymous substitution rates (SSR) for eight pairs of synonymous codons were estimated and the geometric mean of SSR for each toxin genes was calculated. Three toxins (neurotoxin C1, enterotoxins type A and cytotoxin) showed high SSR values. The remaining three toxins fall into the low SSR group. The high CAI and low MD value for neurotoxin C1 and enterotoxins type A can be correlated with their high SSR values. In these two toxins high substitution rate in the synonymous codons may help in the conversion of non-optimal codons to the optimals ones and this explains their high CAI values and low MD values. Horizontally transferred genes are subjected to the mutational processes that also affect the recipient genome (REF). With the progression of the evolutionary time the acquired foreign gene sequences will accumulate substitutions and eventually reflect the DNA composition of the recipient host genome. This process of adjustment of a foreign gene sequence in order to match up the base composition and codon usage of the resident genome is known as 'amelioration''. So the high SSRs in the neurotoxin C1 and enterotoxin type A actually reflect the amelioration process by which the foreign genes gradually adjusted its base composition of the nonoptimal codons and convert it into an optimal one. Relatively low SSR values were found for three toxins (cholera toxin, shiga toxin and diphtheria toxin). These

toxins also showed low CAI values and high mean difference than neurotoxin C1 and enterotoxin type A. For cytotoxin high SSR value was observed which is unusual because it also showed medium CAI value and moderate MD values. In our proposed hypothesis of the conceptual relationship between the evolutionary time and codon usage parameters (CAI, MD, SSR), we assume that the evolutionary time has a linear relationship with CAI and SSR but inverse correlation with MD. We hypothesize that a foreign gene must be subjected a high rate of substitution for a longer period of evolutionary time in order to achieve a high adaptation index. We divide the evolutionary time of acquisition of the toxin genes into three zones. These are distant past, recent past and time span between the distant and recent past. As both the neurotoxin C1 and enterotoxin type A showed high CAI and low MD value they are supposed to reside in the recipient host genome for a relatively longer period of time. The longer period of evolutionary time and high synonymous substitution rate allowed these toxins to adjust its codon usage to according to the host bacteria and this is reflected by their high CAI and low MD values. Inversely when a horizontally transferred gene resides within a host bacterium for a relatively short period of time, it will be subjected to less substitution and consequently it might not be able to achieve optimal codon usage pattern for efficient translation. Similar incident might happen to shiga toxin and diphtheria toxin. They showed relatively low SSR, CAI and high MD value. Consequently we assume that these foreign genes may be acquired in the recent past, as they possess many non optimal codons with respect to their recipient host. Similarly by considering all the parameters we assume that both cholera toxin and cytotoxin were introduced within the time span between distant and recent past. The method that we proposed here is a novel one for of deducing the evolutionary time of acquisition of the foreign genes by prokaryotic genome. Moreover, this method can be applied to any horizontally transferred genes in the prokaryotic genome because the amelioration processes is the same in all prokaryotic genomes but their rate

varies bacteria to bacteria. Here in this method, we utilized the difference of the amelioration rate among different foreign genes to estimate their acquisition time.

Chapter 7

Conclusion

We conclude that, the codon usage analyses of a foreign gene form a basis to estimate their evolutionary time of acquisition by the host bacteria. The uniformity of the amelioration process and codon adaptation of the horizontally transferred genes among different bacteria helps this method to be applied globally to the prokaryotic genomes, which are believed to be assembled by horizontal gene transfer.

Chapter 8

Bibliography

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.

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.

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.

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.

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.

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.

Asher D. Cutter, James D. Wasmuth, and Mark L. Blaxter . The Evolution of Biased Codon and Amino Acid Usage in Nematode Genomes. Mol. Biol. Evol 2006 ; 23(12):2303–2315.

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.

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

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.

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.

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).

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.

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.

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.

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.

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.

Chen SL, Lee W, Hottes AK, Shapiro L, McAdams HH. Codon usage between genomes is constrained by genome-wide mutational processes. Proc Natl Acad Sci U S A. 2004; 9; 101(10):3480-5.

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.

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., 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.

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

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.

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

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.

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.

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.

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.

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

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.

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.

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., 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, 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.

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.

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.

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

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

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

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., (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.

JCat: a novel tool to adapt codon usage of a target gene to its potential expression host Grote A, Hiller K, Scheer M, Mu¨nch R, No¨rtemann B, Hempel DC and Jahn D, Nucleic Acids Research 2005; 33:W526–W531.

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.

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

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

Kolter, R., and C. Yanofsky, (1982). Attenuation in amino acid synthetic operons. Annual Review of Genetics **16:** 113-134.

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.

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

Lawrence JG .Horizontal and Vertical Gene Transfer : The Life History of Pathogens Contrib.Microbiol 2005; 12:255-271

Lawrence JG and Ochman H. Molecular archaeology of the Escherichia coli genome. Proc. Natl. Acad. Sci. USA 1998; 95: 9413–9417

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.

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.

Lloyd, A. T., and P. M. Sharp, (1992a). CODONS - a microcomputer program for codon usage analysis. Journal of Heredity **83:** 239-240.

Maria D. Ermolaeva .Synonymous Codon Usage in Bacteria. Curr. Issues Mol. Biol 2001; 3(4): 91-97.

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.

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.

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.

Moriyama, E. N., and D. L. Hartl, (1993). Codon usage bias and base composition of nuclear genes in *Drosophila*. Genetics **134:** 847-858.

Mouchiroud, D., and C. Gautier, (1990). Codon usage changes and sequence dissimilarity between human and rat. Journal of Molecular Evolution **31:** 81-91.

Muto A, Osawa S. The guanine and cytosine content of genomic DNA and bacterial evolution. Proc Natl Acad Sci U S A 1987 ;84(1):166-9.

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.

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.

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. C. Wilson, (1987b). Evolution in Bacteria: evidence for a universal substitution rate in cellular genomes. Journal of Molecular Evolution **26:** 74-86.

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.

Osawa, S., and T. H. Jukes, (1989). Codon reassignment (codon capture) in evolution. Journal of Molecular Evolution **29:** 271-278.

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.

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.

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.

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.

Pouwels, P. H., and J. A. M. Leunissen, (1994). Divergence in codon usage of *Lactobacillus* species. Nucleic Acids Research **22:** 929-936.

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.

Rosey, E., and G. Stewart, (1989). The nucleotide sequence of the *lacC* and *lacD* genes of *Staphylococcus aureus.* Nucleic Acids Research **17:** 3980.

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 □X174. 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.

Schmidt, W., (1995). Phylogeny reconstruction for protein sequences based on amino acid properties. Journal of Molecular Evolution **41:** 522-530.

Stothard P. JavaScript programs for analyzing and formatting protein and DNA sequences. Biotechniques 2000; 28:1102-1104.

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 PM, Stenico M, Peden JF, Lloyd AT. Codon usage: mutational bias, translational selection, or both? Biochem Soc Trans 1993 ;21(4):835-41.

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

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., 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 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 G. Matassi, (1994). Codon usage and genome evolution. Current Opinions in Genetics and Development **4:** 851-860.

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.

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.

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.

Thioulouse, J., (1990b). Statistical analysis and graphical display of multivariate data on the Macintosh. Computer Applications in the Biosciences **5:** 287-292.

Thompson, K., (1987). Regulation fo sugar transport and metabolism in lactic acid bacteria. FEMS Microbiological Reviews **46:** 221-231.

Walker AEand Bulmer M. Synonymous Substitution Rates in Enterobacteria. Genetics 1995;140:1407-1412

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.

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.

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

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.