


Investigations into the relationship between evolutionary rate variation, protein structure and codon usage

Master Thesis

Author(s):

Du Plessis, Louis 

Publication date:

2011

Permanent link:

<https://doi.org/https://doi.org/10.3929/ethz-a-010380340>

Rights / license:

In Copyright - Non-Commercial Use Permitted



Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Investigations into the relationship between evolutionary rate variation, protein structure and codon usage

A thesis submitted to the
ETH ZÜRICH

for the degree
Master of Science

in
Computational Biology and Bioinformatics

by
Louis du Plessis

supervised by
Maria Anisimova

Computational Biochemistry Research Group
Department of Computer Science
ETH Zürich

July 2011

Abstract

Not all of the processes behind molecular evolution are known. In this report the nonsynonymous and synonymous rates of evolution are examined with respect to the synonymous codon usage and the protein secondary structure in *E. coli K12* and *Homo sapiens*. By applying a model selection step it is first shown that synonymous variation is much more widespread on mammalian genes. It is also shown that there is evidence of selection on the synonymous codon usage in a non-negligible amount of genes, for both organisms. A relationship between the synonymous codon usage and the secondary structure has been reported by several researchers. However, the datasets analyzed here show only a weak relationship. A much stronger relationship is shown to exist between the evolutionary rates and the protein structure. Similarly, a strong relationship exists between preferred and unpreferred codons and the evolutionary rates. These results show that different secondary structures are under different evolutionary pressures. Furthermore, it shows that preferred codons experience different selective constraints to unpreferred codons.

Acknowledgements

First and foremost I would like to thank Maria Anisimova, who suggested the project, supervised me, and was always available to answer my many questions. I would also like to extend my thanks to Sergei Kosakovsky Pond who replied very promptly to all my questions and readily offered explanations and advice. Special thanks also to Melissa Wu and Tamarisk Scholtz who helped out with proofreading the manuscript. Further I would like to thank all my friends and family who have supported me every step of the way. Finally, I would like thank the ETH Zürich Foundation and the committee of the Excellence Scholarship and Opportunity Programme without whose assistance I would not have been able to do my Masters degree at the ETH Zürich.

Contents

1	Introduction	1
2	Causes and measures of unequal codon usage	4
2.1	Measures of codon bias	4
2.1.1	Measuring the usage of individual codons	4
2.1.2	Quantifying the amount of codon bias	5
2.1.3	Identifying preferred and unpreferred codons	5
2.2	Causes of biased codon usage	6
2.3	Relationships between codon usage and protein structure	7
2.3.1	Protein folding and the co-translational hypothesis	7
2.3.2	Translational pauses and rare codon clusters	8
2.3.3	Non-random usage of synonymous codons in secondary structure	9
2.4	Discussion	10
3	Stochastic models of molecular evolution	11
3.1	Calculating the likelihood	12
3.1.1	Independent sites	12
3.1.2	Markovian evolution	12
3.1.3	Independent branches	13
3.2	Codon models	14
3.2.1	Rate matrices	14
3.2.2	Parameter estimation	16
3.3	Models with rate variation	16
3.3.1	Formulation of site-to-site variation	17
3.3.2	Models of site-to-site variation	17
3.4	Models that account for codon usage biases	18
3.5	Discussion	19
4	Implementation and validation of stochastic codon models	20
4.1	Additions to CodonPhyML	21
4.1.1	Implementation of models that allow synonymous rate variation	21
4.1.2	Models that allow selection on codon usage	24
4.1.3	Post-processing	25
4.2	Validation	26
4.2.1	Data	26
4.2.2	Results	26
4.3	Discussion	32
5	Data description and model selection	35
5.1	Background	35
5.1.1	Nested models	36
5.1.2	Non-nested models	36
5.1.3	Sample Size	37
5.2	Data Extraction and Preparation	38
5.2.1	Extraction of orthologous groups	39

5.2.2	Pre-processing of orthologous groups	40
5.2.3	Extraction of preferred codons	41
5.3	Results	42
5.3.1	<i>E. coli</i> dataset	43
5.3.2	Mammalian dataset	53
5.4	Summary	56
6	Relationships between evolutionary rate variation, protein structure and codon usage	58
6.1	Relationships to structure	58
6.1.1	Relationship between amino acid usage, codon usage and secondary structure . .	59
6.1.2	Relationship between evolutionary rates and secondary structure	60
6.2	Correlations between amino acid usage, codon usage and evolutionary rates	64
6.3	Discussion	69
7	Conclusion	70
A	Validation Results	71
B	Extraction of preferred codons	77
C	Comparison of frequency models	82
D	OMA groups with evidence of selection on the synonymous codon usage	90
D.1	<i>E. coli</i> dataset	90
D.2	Mammalian dataset	92
E	OMA groups with a correlation between the nonsynonymous rate and the secondary structure	95
E.1	<i>E. coli K12</i>	95
E.2	<i>Homo sapiens</i>	97
F	OMA groups with a correlation between the synonymous rate and the secondary structure	99
F.1	<i>E. coli K12</i>	99
F.2	<i>Homo sapiens</i>	101
G	Mean rates in secondary structures	103
G.1	<i>E. coli K12</i>	103
G.2	<i>Homo sapiens</i>	104
H	Mean rates in codon classes	105
H.1	<i>E. coli K12</i>	105
H.2	<i>Homo sapiens</i>	106

List of Figures

3.1	An example phylogeny.	14
4.1	Variables stored in memory that are related to the nonsynonymous and synonymous rates of evolution for the three models that were previously implemented in CodonPhyML.	22
4.2	Variables stored in memory that are related to the nonsynonymous and synonymous rates of evolution for the three models that were added to CodonPhyML.	23
4.3	Comparison of the synonymous rates found using CodonPhyML and HyPhy with the Dual GDD and GBDD models coupled with different frequency models.	31
4.4	The performance of the different models of rate heterogeneity as the number of classes are increased.	33
5.1	NCBI phylogenies of the two datasets.	40
5.2	The different models of rate heterogeneity that were used on the <i>E. coli</i> dataset.	43
5.3	Heatmap of the Jaccard indices between the sets of groups selected by different strategies for the Constant Rates and the Dual Γ model on the <i>E. coli</i> dataset.	47
5.4	Histograms of the magnitude of the likelihood increase as the number of classes are increased on the Dual Γ model.	49
5.5	Heatmap of the Jaccard indices between the sets of groups selected by different strategies for the Constant Rates model and models that allow synonymous variation along a GDD on the <i>E. coli</i> dataset.	50
5.6	Histograms of the synonymous and nonsynonymous coefficients of variation on the groups selected for the Dual Γ and GBDD models.	51
5.7	Heatmap of the Jaccard indices between the sets of groups that show selection on the codon usage and the sets of groups selected by forward selection for various models of rate heterogeneity on the <i>E. coli</i> dataset.	52
5.8	Histograms of ω and ψ values estimated with the Constant Rates + ψ model.	52
5.9	Heatmap of the Jaccard indices between the sets of groups that show selection on the codon usage and the sets of groups selected by forward selection on various models of rate heterogeneity on the Mamalian dataset.	55
5.10	The number of groups assigned to each model in the two selection paths that were chosen using forward selection.	56
6.1	Plots of the posterior nonsynonymous and synonymous rates with respect to secondary structures.	61
6.2	The means of the mean posterior rates in the different secondary structures.	63
6.3	Scatter plot of the correlation between the posterior nonsynonymous and synonymous rates over all sites where the respective models were preferred.	65
6.4	Scatter plot of the correlation between the posterior synonymous rate and the RSCU over all sites in all where the respective models were preferred.	66
6.5	The means of the mean posterior rates in the different codon classes.	68
A.1	Comparison of the synonymous rates found using CodonPhyML and HyPhy with the Dual GDD and GBDD models coupled with different frequency models.	76
C.1	Heatmap of the Jaccard indices showing the overlap between groups selected using models on the GBDD and Γ_6 paths coupled with different frequency models	89

List of Tables

4.1	Statistics of the five datasets used for validation.	26
4.2	Validation results for the Constant Rates, Nonsynonymous Γ and Nonsynonymous GDD models.	28
4.3	Validation results for the Dual Γ and Dual GDD models.	29
4.4	Validation results for the GBDD model.	30
4.5	Validation results for the Constant Rates + ψ model.	32
5.1	Assigning support to models using the AIC.	37
5.2	Statistics of the two datasets.	41
5.3	Preferred codons selected for the two datasets.	42
5.4	The results of using information theoretic criteria to select between the 10 candidate models of rate heterogeneity on the <i>E. coli</i> dataset.	44
5.5	Results of the different selection strategies on the <i>E. coli</i> dataset for models where rate heterogeneity is treated with a Γ distribution, for different numbers of classes.	46
5.6	Results of the different selection strategies on the <i>E. coli</i> dataset for models where rate heterogeneity is treated with a GDD distribution.	48
5.7	The number of groups in the <i>E. coli</i> dataset showing significant improvement when the ψ parameter is added to different models.	53
5.8	Results of the different selection strategies on the Mammalian dataset for the two model selection paths that were used to model rate heterogeneity.	54
5.9	The number of groups in the Mammalian dataset showing significant improvement when the ψ parameter is added for different models.	55
6.1	The numbers of groups assigned to site-specific datasets used to search for pairwise correlations in this chapter.	59
6.2	Codons that show a significant deviation from their expected usages within their synonymous codon families. The datasets are as defined in table 6.1.	60
6.3	The number of sequences that show a significant bias between the posterior evolutionary rates and the secondary structure. The datasets are as defined in table 6.1.	62
6.4	The number of sequences that show a significant bias between the posterior evolutionary rates and the codon usage. The datasets are as defined in table 6.1.	67
A.1	Validation results for the Constant Rates, Nonsynonymous Γ and Nonsynonymous GDD models.	72
A.2	Validation results for the Dual Γ and Dual GDD models.	73
A.3	Validation results for the GBDD model.	74
A.4	Validation results for the Constant Rates + ψ model.	75
B.1	Preferred codons selected for <i>Homo sapiens</i> and <i>E. coli K12</i> as the selection criteria are varied.	78
B.2	Comparison between the codons identified in other publications and the method described here.	79
B.3	Preferred codons selected for the Mammalian dataset	80
B.4	Preferred codons selected for the <i>E. coli</i> dataset.	81

C.1	Results of the different selection strategies on the <i>E. coli</i> dataset for models where rate heterogeneity is treated with a Γ distribution, for different numbers of classes.	83
C.2	Results of the different selection strategies on the <i>E. coli</i> dataset for models where rate heterogeneity is treated with a Γ distribution, for different numbers of classes.	84
C.3	Results of the different selection strategies on the <i>E. coli</i> dataset for models where rate heterogeneity is treated with a GDD distribution.	85
C.4	Results of the different selection strategies on the <i>E. coli</i> dataset for models where rate heterogeneity is treated with a GDD distribution.	86
C.5	Results of the different selection strategies on the Mammalian dataset for the two model selection paths that were used to model rate heterogeneity.	87
C.6	Results of the different selection strategies on the Mammalian dataset for the two model selection paths that were used to model rate heterogeneity.	88
D.1	OMA groups in the <i>E. coli</i> dataset where positive selection on the codon usage was detected.	90
D.2	OMA groups in the <i>E. coli</i> dataset where negative selection on the codon usage was detected.	91
D.3	OMA groups in the Mammalian dataset where positive selection on the codon usage was detected.	92
D.3	OMA groups in the Mammalian dataset where positive selection on the codon usage was detected (continued).	93
D.3	OMA groups in the Mammalian dataset where positive selection on the codon usage was detected (continued).	94
D.4	OMA groups in the Mammalian dataset where negative selection on the codon usage was detected.	94
E.1	OMA groups for <i>E. coli</i> sequences from the GBDD ($+\psi$) dataset where a correlation between the posterior nonsynonymous rate and the secondary structure was detected.	95
E.2	OMA groups for <i>E. coli</i> sequences from the Dual Γ ($+\psi$) dataset where a correlation between the posterior nonsynonymous rate and the secondary structure was detected.	96
E.3	OMA groups for <i>Homo sapiens</i> sequences from the GBDD ($+\psi$) dataset where a correlation between the posterior nonsynonymous rate and the secondary structure was detected.	97
E.4	OMA groups for <i>Homo sapiens</i> sequences from the Dual Γ ($+\psi$) dataset where a correlation between the posterior nonsynonymous rate and the secondary structure was detected.	98
F.1	OMA groups for <i>E. coli</i> sequences from the GBDD ($+\psi$) dataset where a correlation between the posterior synonymous rate and the secondary structure was detected.	99
F.2	OMA groups for <i>E. coli</i> sequences from the Dual Γ ($+\psi$) dataset where a correlation between the posterior synonymous rate and the secondary structure was detected.	100
F.3	OMA groups for <i>Homo sapiens</i> sequences from the GBDD ($+\psi$) dataset where a correlation between the posterior synonymous rate and the secondary structure was detected.	101
F.4	OMA groups for <i>Homo sapiens</i> sequences from the Dual Γ ($+\psi$) dataset where a correlation between the posterior synonymous rate and the secondary structure was detected.	102
G.1	The means of the mean posterior rates in the different secondary structures, for <i>E. coli K12</i>	103
G.2	The means of the mean posterior rates in the different secondary structures for <i>Homo sapiens</i>	104
H.1	The means of the mean posterior rates in the different codon classes, for <i>E. coli K12</i>	105
H.2	The means of the mean posterior rates in the different codon classes, for <i>Homo sapiens</i>	106

Chapter 1

Introduction

All of the necessary information for defining an organism is stored in its DNA sequence¹, consisting of a double-stranded chain of four different nucleotides. Genes are regions of the genome that code for proteins, which are used to perform cellular functions and make up most of the structures in a cell. Genes are first transcribed into single-stranded mRNA sequences by RNA polymerases, before being translated into proteins by ribosomes. This process is known as the central dogma of molecular biology. While there is a one-to-one correspondence between DNA and RNA, this is not the case with proteins. Where DNA and RNA are composed of four different types of nucleotides, proteins are composed of 20 different amino acids. During translation, every trinucleotide (called a codon) codes for one amino-acid. This is known as the genetic code. There are 64 different combinations of three nucleotides each, which means that the genetic code is redundant. Of the 64 possible combinations, 61 code for amino acids in the standard genetic code. The remaining three codons are used to signal the end of a gene (stop codons). Of course, the process is much more complicated than this simple explanation and involves several other steps. However, these steps are not necessary for an understanding of the research presented here, and details may be found in any standard molecular biology textbook, for instance [Alberts *et al.* 2007].

A translated amino-acid chain folds into a complex three-dimensional structure. The function of a protein is defined by its structure. Although it was initially thought that the amino acid sequence carries all the information needed to dictate a protein's structure [Anfinsen 1973], attempts at performing structure prediction have been unsuccessful thus far. Over the course of time several other factors have been identified that can have an impact on a protein's conformation. The research presented here pays special attention to the possibility that the mRNA sequence plays an important role in deciding a protein's final conformation. Protein structure can be viewed at three different resolutions. The primary structure of a protein is its amino acid sequence. Secondary structures are regular structural elements interspersed in a protein structure. The tertiary structure of a protein denotes its precise three-dimensional structure. Some researchers add a fourth type of structure, quaternary structure, to denote different, largely independent, protein domains that are connected by unstructured chains. Proteins are commonly partitioned into three types of secondary structures, helices, sheets and loops. While helices are constructed from neighbouring amino acids, sheets are formed by distant strands packing together. Loops (or coils) are short chains of amino acids that connect the other elements. The most common secondary structures are alpha helices and beta sheets.

Processes in a cell are not completely error-free and mistakes may be introduced in the DNA sequence during the replication process. If one of these random mutations fall within a gene it could have an effect on the function of the protein that the gene codes for. Even a change of just one nucleotide can alter the structure or function of a protein, which may in turn cause a change in the organism's phenotype. The vast majority of mutations will be harmful to the organism and will not be propagated to future generations. Even if a harmful mutation is propagated, it is unlikely that it will become fixated in the species. Occasionally a mutation may increase the fitness of a gene, possibly by making it more stable, adding a novel function or simply by optimizing its translation. Due to the effects that mutations have on fitness the molecular evolution of a gene is under selective pressures to allow only those mutations that are beneficial. Different sites may be under different selective

¹Or RNA sequence, in the case of certain viruses.

pressures. If a site is under pressure to remain the same and not allow any mutations, that site is said to be under purifying (or negative) selection. On the other hand, if mutations at a site are usually beneficial, then that site is under diversifying (or positive) selection. Lastly, a mutation may also be neutral and cause no perceivable change in a gene’s fitness. As an example, loops are usually more mutable than helices and sheets, hence it is thought that they are under less pressure from purifying selection than the other structural categories.

Due to the degeneracy of the genetic code, two different types of mutations are possible. If a mutation changes the amino acid that a codon codes for it is called a nonsynonymous mutation. On the other hand, synonymous mutations have no effect on the amino acid chain of the translated protein. Originally it was thought that only nonsynonymous mutations can have an effect on the fitness of a protein. This led to the assumption that all synonymous mutations are accepted, and hence that the synonymous rate of mutation is equal to the neutral rate of mutation, due to the lack of selective pressures [Chamary *et al.* 2006]. However, this belief has changed in recent years and it is now thought that synonymous mutations also play an important role in a gene’s fitness. This is most obviously seen in the phenomenon of codon usage biases within genes. If all synonymous codons are equally fit, they should all be used with equal frequency. However, this is not what is seen in practice. Although the majority of codon usage biases can be attributed to random mutational preferences for certain nucleotides, this cannot be used to describe all of the observed biases. It has been theorized that synonymous codon choices may cause translational pauses, and that these pauses aid the folding process. Although several studies have shown the existence of an effect, their results are often incompatible with each other. Moreover, the extent of the effect is not known. It is possible that the effect of synonymous codon choices only increases the translation speed and probability of correct folding, and are not essential for correct folding [Komar 2009, Oresic and Shalloway 1998]. Regardless of the causes behind selection on synonymous codons, it is clear that a selective constraint should be visible as a change in the synonymous rate of evolution [Shields *et al.* 1988]. However, it has been argued that the selective constraints on synonymous codons are very weak and that synonymous substitutions are dominated by random mutations. If this is the case it would be very difficult to detect an effect. Some causes of codon biases and possible selective constraints on synonymous codons are discussed in chapter 2. In particular, the chapter focusses on the relationship between the synonymous codon usage and the protein structure.

The aim of this research is to investigate correlations between evolutionary rate variation, synonymous codon usage bias and protein secondary structure. If there is a relationship between the codon usages and the secondary structures, then it should also be possible to detect a relationship between synonymous rate and different secondary structural classes. A relationship between the nonsynonymous rate and the secondary structure may also be assumed, due to the different evolutionary pressures on these structures. This study was performed on two large datasets, extracted from the PDB and OMA databases (see section 5.2 for a description of the data extraction procedure). One dataset is based around Mammalians, while the other consists of various *E. coli* strains. Using these datasets correlations between the evolutionary rates, codon usages and secondary structures were investigated for *Homo sapiens* and *E. coli K12*.

In order to obtain the nonsynonymous and synonymous rates at every codon a stochastic model of evolution was fitted to each gene in the datasets. Stochastic models of evolution are discussed in detail in chapter 3. The models used here are of a class of models known as random effects likelihood models, that find evolutionary parameters for a set of orthologs by maximum-likelihood. In the models described here several simplifying assumptions are made. The three most important assumptions made are that evolution proceeds independently on different codons and lineages and that the evolution of a sequence is not dependent on its history. All of these assumptions are falsifiable, however more complicated models are more difficult to optimize and hence the simulation becomes intractable. CodonPhyML was used to perform the simulations [Zanetti 2010]. CodonPhyML’s repertoire of models was also expanded by adding models that allow synonymous rate variation and models that allow selection on the codon usage. All of the parametric codon models in CodonPhyML were then validated on a separate dataset by comparing the results to HyPhy, a similar package. Although there are differences in the results returned by some of the more complicated models, these differences can be ascribed to the complexity of the likelihood surfaces of these models. The changes made to CodonPhyML, as well as the validation experiments are detailed in chapter 4.

By fitting different models of varying complexity to the datasets and employing a model selection

strategy, genes that show a significant variation in the synonymous rate and genes that have evidence of the presence of selection for codon usage were identified. Among the *E. coli* strains, around one fifth of the genes show a significant variation in the synonymous rate. The number of genes showing selection on the codon usage is a little smaller. In contrast, on the Mammalian dataset most genes show evidence of synonymous variation. The number of genes with evidence of selection on the codon usage is less, with about a third of the genes showing a significant result. These findings, as well as the model selection strategy followed, are given in chapter 5.

Finally, correlations between the evolutionary rates, codon usages and secondary structures are examined in chapter 6. Although only a very weak effect is visible between the synonymous codon usage and the secondary structural classes, the effects between both the nonsynonymous and synonymous rates and the structural classes are much stronger in both organisms. However, this effect is not directly related to the codon usage within a gene, as there does not exist a strong correlation between the synonymous codon usage and the evolutionary rates. Instead, a strong effect is observable between the evolutionary rates and their usages in preferred or unpreferred codons. These results indicate that there is indeed a significant relationship between the evolutionary rates and secondary structures. Furthermore, there is also a strong relationship between the evolutionary rates and the synonymous codon choice, although this is not directly influenced by the codon usage within any particular gene, but instead by the codon usage within an organism.

Chapter 2

Causes and measures of unequal codon usage

Biases in the synonymous codon usage is a universal phenomenon and have long been observed in both prokaryotes and eukaryotes [Comeron and Aguad 1998]. However, the amount of bias and the codons that are biased, vary between organisms and even between genes within one organism.

When measuring the codon bias, a distinction should be made between the bias observed for an individual codon in a sequence, and the overall amount of bias in the sequence. The bias of an individual codon can serve to identify rare or abundant codons while the overall amount of bias in a sequence can be used to identify highly expressed genes, since it is known that such genes have a highly biased codon usage and sometimes make use of only a small subset of preferred codons [Gupta *et al.* 2000, Thanaraj and Argos 1996a]. These measures, as well as some strategies for identifying preferred codons are introduced in the next section. In section 2.2 some of the proposed causes for codon bias are discussed. Section 2.3 examines the hypothesis that the codon usage may be related to the protein structure.

2.1 Measures of codon bias

2.1.1 Measuring the usage of individual codons

When measuring the codon usage it is important to take the amino acid bias into account. Not all amino acids are used with equal frequencies in coding sequences, hence using the absolute frequency of a codon in a sequence is not a good measure of the codon preference. A better measure is the synonymous codon frequency. For an amino acid, aa_i , its synonymous codon family is defined as $SC(aa_i) = \{sc_j | t(sc_j) = aa_i\}$, where $t(x)$ is the amino acid that codon x codes for in the genetic code. Define n_i as the number of synonymous codons that code for the amino acid, that is $n_i = |SC(aa_i)|$. Further, define $obs(aa_i, sc_j)$ as the number of occurrences of codon sc_j for amino acid aa_i in the sequence of interest. The synonymous codon frequency is given by:

$$f_{syn}(sc_j) = \frac{obs(aa_i, sc_j)}{\sum_{k=1}^{n_i} obs(aa_i, sc_k)}$$

The synonymous codon frequency measures the frequency of a codon in its synonymous codon family. This is a good measure when different amino acids are treated independently. However, this measure cannot be used in comparisons between amino acids or genes, as different amino acids have different levels of redundancy. As an example, suppose that the codons CTT and AAG both have synonymous codon frequencies of 0.5. Although the values are identical, the synonymous codon frequency of CTT is much more significant, as it codes for Arginine, which has 6-fold redundancy, whereas AAG codes for Lysine, which has only 2-fold redundancy. Hence, the number of codons in a synonymous codon family should be accounted for. This is done with the relative synonymous codon frequency:

$$RSCU(sc_j) = n_i f_{syn}(sc_j) = \frac{obs(aa_i, sc_j) / \sum_{k=1}^{n_i} obs(aa_i, sc_k)}{1/n_i}$$

which removes most of the bias due to the amino acid usage [Peden 1999].

2.1.2 Quantifying the amount of codon bias

The measures discussed in this section attempts to give a summary statistic of the amount of codon bias observed in a sequence. They can be grouped into two classes, measures that compare the observed frequency of a synonymous codon to the frequency of preferred codons, and measures that assume as a null hypothesis that all synonymous codons are used equally [Comeron and Aguad 1998]. Clearly, measures of the first kind require *a priori* information on which codons are preferred. Only two measures, \hat{N}_c and \hat{N}'_c are discussed in detail. For a review of other methods see [Peden 1999].

The most widely used measure of the first kind is the codon adaptation index (CAI). This measure compares the frequency of a codon's usage to its usage in a reference set of highly expressed genes [Comeron and Aguad 1998, Thanaraj and Argos 1996b].

Among measures of the second kind, the scaled χ^2 measure uses the χ^2 -test statistic on equal usage of synonymous codons, divided by the number of codons in the sequence [Shields *et al.* 1988]. Using simulations, Comeron and Aguad [1998] found that the sequence length affects the scaled χ^2 measure, with the value being overestimated for long sequences.

The effective number of codons (\hat{N}_c) is derived from the expected number of alleles at a locus, n_e . Treating amino acids as loci and synonymous codon choices as alleles, Wright [1990] constructed \hat{N}_c as follows. The homozygosity of codon usage for amino acid aa_i is given by:

$$\hat{F}_{aa_i} = \left(n_{aa_i} \sum_{j=1}^{n_i} f_{syn}(sc_j)^2 - 1 \right) / \left(n_{aa_i} - 1 \right)$$

where n_{aa_i} is the number of occurrences of the amino acid in the sequence. The expected number of alleles may be estimated by $1/\hat{F}$. For each r -fold redundancy class the mean value of \hat{F} , \hat{F}_r is calculated. The value of \hat{N}_c is then given by the following sum:

$$\hat{N}_c = 2 + \frac{9}{\hat{F}_2} + \frac{1}{\hat{F}_3} + \frac{5}{\hat{F}_4} + \frac{3}{\hat{F}_6} \quad (2.1)$$

for the standard genetic code. Comeron and Aguad [1998] found \hat{N}_c to be the most well-behaved measure of the second kind. However, \hat{N}_c measures the departure from random usage and does not take nucleotide biases into account. Hence, biases reported with \hat{N}_c may only be due to the background nucleotide composition of the sequences. A correction to \hat{N}_c , \hat{N}'_c , accounts for nucleotide biases by measuring the departure from an expected distribution derived from the background nucleotide composition [Novembre 2002]. The only modification to the calculation is that the equation for \hat{F} is modified to:

$$\hat{F}_{aa_i} = \frac{X_{aa_i}^2 + n_{aa_i} - n_i}{n_i(n_{aa_i} - 1)}$$

where $X_{aa_i}^2$ is the χ^2 statistic measuring the departure of the amino acid from its expected usage, defined as:

$$X_{aa_i}^2 = \sum_{j=1}^{n_i} \frac{n_{aa_i} \left(f_{syn}(sc_j) - f_{exp}(sc_j) \right)^2}{f_{exp}(sc_j)} \quad (2.2)$$

where $f_{exp}(sc_j)$ is the expected frequency of the codon. Novembre [2002] found that \hat{N}'_c is still unbiased with respect to the sequence length.

2.1.3 Identifying preferred and unpreferred codons

It is clear that preferred synonymous codons will occur at a higher frequency than unpreferred (or rare) codons. A very simple measure to identify preferred or unpreferred codons is to use the RSCU.

For instance, Thanaraj and Argos [1996b] labelled all codons with an RSCU of less than 0.08 as rare. Of course, this measure does not account for the background nucleotide composition at all and the cutoff value is arbitrarily chosen. A more robust method is to use a statistical test to find codons that show a significant increase or decrease of usage in highly expressed genes.

Highly expressed genes may be identified computationally by one of the summary statistics introduced in the previous section. Along this train of thought Zhou *et al.* [2010] used \hat{N}'_c , while Akashi [1995] used the scaled χ^2 measure. Ingvarsson [2008] used an alternative approach, where the number of EST sequences that are matched to a gene is used as a proxy for the level of expression.

Wu *et al.* [2010] used a different approach altogether to identify rare codons. A statistical test based on random usage is used to identify rare codon pairs. Rare codons are subsequently identified by finding the codons that make the largest contributions to the rare codon pairs, as assessed by a hypergeometric test.

2.2 Causes of biased codon usage

A number of explanations have been proposed to explain the observed codon biases. For a recent review see [Plotkin and Kudla 2011].

It was traditionally assumed that synonymous mutations are neutral and do not cause any change in a gene's fitness [Chamary *et al.* 2006]. If this assumption holds, the synonymous substitution rate is proportional to the point mutation rate [Parmley *et al.* 2006]. If the neutral theory is true, codon biases may be explained by mutational effects, such as specific biases in nucleotide mutations, contextual biases in the point mutation rates and biases in repair [Plotkin and Kudla 2011]. Mutational biases such as these are responsible for maintaining the background nucleotide content of a genome. Moreover, most vertebrate genomes are divided into isochores, large stretches with different nucleotide compositions, also thought to be the effect of mutational biases [Comeron and Aguad 1998]. The GC content of a genome, (or isochores) can be used to explain the majority of the observed codon bias [Shields *et al.* 1988]. In fact, the differences in codon bias between bacterial genomes can be predicted by their respective GC contents [Plotkin and Kudla 2011]. However, mutational biases cannot explain all of the observed codon biases [Hershberg and Petrov 2008, Urrutia and Hurst 2001, Oresic and Shalloway 1998]. In particular, mutational biases are caused by genome-wide processes, which cannot be used to explain the variation between genes in the same genome. Furthermore, a direct correlation between the codon bias and the corresponding tRNA abundance has been measured, which can also not be explained by a mutational bias [Hershberg and Petrov 2008].

To explain the correlation between the tRNA abundance and the codon bias it has been postulated that natural selection acts on the synonymous codon choice to increase the accuracy and speed of translation [Hershberg and Petrov 2008, Ingvarsson 2008, Comeron and Aguad 1998]. Supporting evidence for this postulate includes the fact that the codon bias is related to the level of expression and the fact that highly expressed genes preferentially use codons with the most abundant tRNAs [Urrutia and Hurst 2001, Comeron and Aguad 1998, Shields *et al.* 1988]. It has also been observed that the expression of transgenes can be increased dramatically by replacing rare codons [Plotkin and Kudla 2011, Urrutia and Hurst 2001]. Furthermore, it has been found that the most accurately translated synonymous codons are preferred at functionally important sites [Akashi 1994]. Finally, Ingvarsson [2008] found evidence of selection on synonymous codons in plants. However, there is only a weak correlation between codon bias and the expression level in mammals, and even some cases where genes with a high codon bias are not highly expressed [Chamary and Hurst 2005b]. It should also be noted that selection for the speed of translation will only be effective if elongation, and not initiation of translation is the rate-limiting step [Plotkin and Kudla 2011]. Nevertheless, selection for translational efficiency appears to be the largest cause of selection on the codon usage. In these cases synonymous substitutions are not neutral, but under diversifying (or positive) selection for mutations to preferred codons, and under purifying (or negative) selection against mutations away from preferred codons.

Selection for specific synonymous codons may also be due to other effects. Especially in vertebrates, certain nucleotides are preferred near splice sites or other motifs, such as exonic splicing enhancers (ESEs) and this could lead to a codon bias in these regions [Parmley and Hurst 2007, Parmley *et al.* 2006, Chamary and Hurst 2005a]. The mRNA secondary structure is also affected by synonymous

codon choices and it has been shown that stable structures have a higher fitness [Chamary and Hurst 2005b]. Synonymous codons may also be selected for translational pauses, which may have an effect on protein folding. (Translational pauses are examined in detail in the next section). Several other effects, such as post-translational modifications and nucleosome positioning may also be related to the synonymous codon usage [Plotkin and Kudla 2011]. Finally, several cases are known of diseases that are caused by synonymous mutations [Chamary *et al.* 2006, Chamary and Hurst 2005b]. In all of these cases purifying selection is expected to act on the sites in question, to avoid mutations to less favourable codons. Unpreferred codons may also be under purifying selection here, because these sites do not cover a large proportion of the sequence. Lastly, it is still possible for synonymous mutations to be under selection without there being any clear codon preference [Chamary and Hurst 2005b].

Whatever the case may be, it is believed that the selective pressures on synonymous codons are very weak, which would explain why most of the variation can be attributed to mutational effects. However, selection will be effective as long as $N_e s > 1$, where N_e is the effective population size and the s is the selection coefficient, which is equal to the difference in fitness between the synonymous codons [Shields *et al.* 1988]. Hence, selection on codon usage is thought to be more pronounced for bacteria and other organisms with a large effective population size. On the other hand, small population sizes mean that the effect of selection on synonymous sites is too small to lead to fixation, which may make it harder to observe the effects of selection on vertebrates [Chamary and Hurst 2005b].

2.3 Relationships between codon usage and protein structure

2.3.1 Protein folding and the co-translational hypothesis

Two extremely difficult problems are associated with protein folding. The first is to determine the three-dimensional structure of a protein in its native state. The second is to investigate the folding pathway by tracking the conformations followed by the protein whilst transitioning between its denatured and native states.

The three dimensional structure may be investigated experimentally using X-ray crystallography, nuclear magnetic resonance and a host of other spectral methods. However, these methods are slow and not applicable to all proteins. Computational structure prediction usually relies on machine learning approaches such as neural networks (see for instance [Brunak and Engelbrecht 1996]), but even the most advanced prediction algorithms are still unreliable.

Predicting the structure of a protein is a subproblem of determining the folding pathway. Naturally, if the folding pathway is known, the final conformation is also known. The folding pathway is usually investigated by a combination of experimental and computational approaches. Experimentally, proteins are denatured (either thermally or chemically) and their conformations are studied by spectral methods. Computationally, molecular simulations may be used to gain information about the interactions between the molecules that constitute a protein. However, current technology only allows simulations on small peptides over short intervals.

Originally, it was assumed that after translation, a protein examines random conformations until it finds its natural state, which is the conformation with the smallest free energy. Levinthal raised the paradox that if this were true even small proteins would take longer than the age of the universe to fold [Daggett and Fersht 2003]. Nowadays most people believe that proteins fold along defined pathways that restrict the number of possible conformations that may be examined [Daggett and Fersht 2003]. Two different theories have been proposed. The framework model states that structure forms in a stepwise manner, initially secondary structures form, and then the different secondary structural elements pack together into protein domains. On the other hand, the hydrophobic-collapse model states that folding is initiated by a hydrophobic collapse which forms the core of the protein around which other parts fold.

The framework model is supported by the co-translational folding hypothesis, which was first proposed by Purvis *et al.* [1987]. Previously, it was believed that mRNA strands were completely translated by ribosomes before folding began. Predictions of the folding pathway only concentrated on the spatial separation of amino acids. Based on several observations, Purvis *et al.* [1987] argued that translation and folding are not separate events, but happen at the same time. This observation is motivated by the fact that translation and folding do not happen on separate timescales. In some cases

the time taken for translation may even be longer than the time it takes the same protein to refold *in vitro*. Hence, the authors argue that the translated polypeptide chain starts to fold into conformations before translation is finished. This is supported by evidence that shows that the ribosome tunnel is large enough to accommodate helices and even chain to chain interactions [Komar 2009, Oresic and Shalloway 1998, Thanaraj and Argos 1996a]. If folding starts before translation is complete the amino acids that are available for interactions are limited, thereby restricting the folding pathway. Although there is evidence that some proteins only start folding after translation, several experimental results also support the co-translational folding hypothesis [Komar 2009].

It is obvious that the rate of elongation plays a role in the pathway followed during co-translational folding. Purvis *et al.* [1987] further argue that the rate of elongation varies and that there is evidence of translational pauses. They argue that translational pauses can provide a selective advantage in giving the translated portion of the polypeptide enough time to find its native conformation before translation recommences. The authors propose that translational pauses may be present between protein domains and secondary structural units to aid in speeding up the folding process, especially in proteins with a difficult folding pathway.

Anfinsen's hypothesis claims that all the information needed for a protein to fold correctly is contained in its amino acid sequence [Anfinsen 1973]. While this does seem to be the case for some proteins (in particular small, globular proteins), it is often difficult to refold proteins *in vitro* [Komar 2009]. There is evidence to show that protein-folding *in vitro* is a lot less efficient than folding *in vivo*, which is an indication that factors beyond the amino acid sequence may be involved in the folding process. It is known that there are several chaperone and catalyst proteins that assist in the folding process, however, it has been argued that these proteins have a bigger effect on the yield of a protein than on the folding mechanism itself [Komar 2009]. The co-translational hypothesis suggests that the mRNA sequence may add an additional layer of information to the peptide sequence, by modulating the rate of translation.

2.3.2 Translational pauses and rare codon clusters

It is believed that the rate of elongation is primarily affected by the non-random usage of synonymous codons [Komar 2009]. As stated previously, not all tRNA species occur with the same frequency, and this is thought to be the driving factor behind the rate of translation, and hence a major cause of codon bias (see section 2.2). Purvis *et al.* [1987] argue that rare codon clusters may be responsible for translational pauses between protein structural units, which could aid proteins in finding their native conformations. Depending on the study, the structural units have been defined as protein domains or as secondary structural elements [Brunak and Engelbrecht 1996].

The presence of translational pauses may be investigated by looking at the sizes of nascent polypeptides. By using codon frequency profiles Krasheninnikov *et al.* [1991] and Guisez *et al.* [1993] showed that rare codon clusters may be responsible for translational pauses.

Several experimental studies have been performed where researchers replaced rare codons with non-rare codons, and subsequently showed that the modified protein was translated faster, but showed reduced activity [Crombie *et al.* 1992, Komar *et al.* 1999, Cortazzo *et al.* 2002]. This is used as indirect evidence that the modified proteins are prone to misfolding. In contrast, it has also been shown in at least one example that introducing rare codons when expressing genes in a foreign organism can increase the yield of the protein [Komar 2009]. This is thought to be an effect of emulating the mRNA translation kinetics of the gene in the foreign organism.

A systematic search for rare codon clusters was performed by Brunak and Engelbrecht [1996]. However, no significant stretches of rare codons were found. Moreover, only one of the rare codon clusters they found is located between structural domains, while the rest of the clusters are within alpha helices or beta sheets. In fact, they found that the termini of secondary structural elements are dominated by non-rare codons. They did find a correlation between the mRNA sequence and the protein structure, but this is ascribed to the preferential usage of certain amino acids at the ends of structural elements.

Makhoul and Trifonov [2002] searched for rare codon clusters by looking for minima of the codon bias in codon frequency profiles. The authors report two minima for prokaryotic sequences, one at the start of the protein, and one after about 150 codons. They only ascribe the second minimum to a translational pause for co-translational folding, and note that most prokaryotic protein domains are

around 150 codons in length.

By using the codon frequency as a proxy for the elongation rate Thanaraj and Argos [1996b] found that slow regions of the mRNA chain preferentially code for domain boundaries in *E. coli*. Extending this result led to the finding that alpha helices are encoded by translationally fast regions and beta strands by slow regions [Thanaraj and Argos 1996a]. However, their dataset was small and it has been pointed out that they made use of the absolute codon frequency, which could have introduced a second-order correlation to the amino acid usage [Oresic and Shalloway 1998].

Despite evidence to the contrary, several examples are known where rare codon clusters occur between structural domains of proteins [Komar 2009, Komar and Jaenicke 1995]. Furthermore, the location of rare codon clusters seems to be conserved between homologs, which may be an indication of the importance of translational pauses. Overall, these results seem to indicate that rare codon clusters are important for the correct folding of a protein. However, it should be noted that translational pauses may also be caused by a number of other factors, such as the mRNA secondary structure, context effects from neighbouring codons and codon-anticodon interactions [Thanaraj and Argos 1996b, Guisez *et al.* 1993].

2.3.3 Non-random usage of synonymous codons in secondary structure

Several researchers have looked for evidence to support the co-translational hypothesis by studying the slightly more general problem of identifying differential codon usages in different secondary structural types. This is usually done by assuming random use of synonymous codons in the different structure classes and then using a statistical test to evaluate the bias in codon usage. Even when an effect is found, it may not be significant, as second order correlations could easily be introduced if one is not careful [Oresic and Shalloway 1998].

Biases can be easily introduced when computing the expected frequency of a synonymous codon within different structural types. As discussed in section 2.1.1, certain codons are preferred in each synonymous codon family. This bias is not only dependent on the organism, but also on the location of the gene within the genome. When testing for the biased codon usage within different secondary structures this bias needs to be accounted for. Another important bias that should be accounted for is the amino acid bias, since it is well known that some amino acids have a higher propensity for particular secondary structures [Chou and Fasman 1974]. Adzhubei *et al.* [1996] measured the expected number of a synonymous codon, sc_i , within a structural class, ss_j as:

$$\text{obs}_{exp}(sc_i, ss_j) = \text{obs}_{aa}(sc_i, ss_j) f_{syn}(sc_i)$$

where obs_{aa} is the number of times that the amino acid that sc_i codes for occurs in ss_i and f_{syn} is defined as in section 2.1.1. On the other hand, Mukhopadhyay *et al.* [2007] and Gupta *et al.* [2000] only accounted for the nucleotide bias. Correcting only for the nucleotide bias within different secondary structural types assumes that the nucleotide compositions of secondary structures are conserved. This does not completely take the amino acid or synonymous codon biases of a gene into account. Correcting for the amino acid usage is more important, since correcting for the nucleotide usage alone would not completely remove the amino acid bias. Although effects between the nucleotides in neighbouring codons may cause a correlation to the structure [Oresic and Shalloway 1998], it is not immediately clear how to remove this bias. It is assumed that any further biases between the nucleotide composition and the secondary structure are minor and can be ignored. Other researchers have avoided the explicit calculation of expected frequencies by using Monte Carlo simulations [Oresic and Shalloway 1998], or by shuffling sequences to obtain the expected distributions [Tao and Dafu 1998].

Statistical tests are usually performed independently for each synonymous codon family, or in some cases for each codon. Oresic and Shalloway [1998] argues that a multiple test correction should be applied. They computed p -values by Monte Carlo simulations to give the probability of observing an effect in any of the contingency tables. However, they detected no effect between the codon usage and the secondary structure. Oresic and Shalloway [1998] further argue that the synonymous codon used at a site may have an effect on the secondary structures a few codons from it. To test this hypothesis they measured the correlation between a codon and the secondary structure occurring within a 10 amino acid radius from it. In this case they detected a significant bias on one synonymous codon each from *E. coli* and *Homo sapiens*. However, Komar [2009] states that although this hypothesis is reasonable, it is usually observed that rare codons are correlated to the structure at the codon itself.

Due to the large assortment of different methods and datasets that have been employed to investigate effects between the codon usage and secondary structure the results from different studies are rarely compatible. For instance Tao and Dafu [1998] found an effect for *Homo sapiens*, but no effect for *E. coli*. In contrast to this, Gu *et al.* [2004] reports an effect for *E. coli*, but no effect for *Homo sapiens*. Nevertheless, it seems clear that correlations between the codon usage and secondary structures do exist in both prokaryotes and eukaryotes. It also seems clear that the correlation is specific to the species [Gupta *et al.* 2000, Oresic and Shalloway 1998]. Furthermore, the large observed difference between prokaryotes and eukaryotes could mean that the correlations are caused by different factors.

2.4 Discussion

In this chapter, measures and causes of codon biases are discussed. The presence of selection on synonymous codon usage implies that synonymous sites are not neutral as previously believed. This raises the possibility of correlations between the synonymous substitution rate and the codon usage. In particular, the synonymous rate is expected to be low for sites under negative selection. Such sites may be synonymous sites with preferred codons, or rare codons at the sites of translational pauses and other conserved motifs. Alternatively, if selection for translational efficiency is assumed, sites with unpreferred codons would be under positive selection and show high synonymous rates. However, the effect of selection on the synonymous codon usage is expected to be very small, thus any correlation would be difficult to observe. Further, it is also expected that most synonymous sites will be neutral and not under any selective pressures [Chamary *et al.* 2006]. Urrutia and Hurst [2001] found a weak inverse correlation between the average synonymous rate and the codon bias. However, they ascribe this bias to the sequence length and not to selection.

A possible problem with relating synonymous codon choices to secondary structures is that the denatured state of a protein is not devoid of structure [Daggett and Fersht 2003]. Unfolded protein chains do show some structures, but not the same structures as proteins in their natural states. While it is likely that the polypeptide may start folding before its translation is finished, the structure it assumes may not be its final structure. In fact, for many proteins intermediate structures have been identified between their denatured and natural states [Daggett and Fersht 2003]. Hence, even if a correlation exists between the codon usage and the protein structure, it may not be possible to show this correlation by comparing the final structure and the codon usage.

Chapter 3

Stochastic models of molecular evolution

Stochastic models of molecular evolution may be used to shed light on the evolutionary process. In particular, models may be used to gain knowledge on the evolutionary relationships between sequences, their divergence times and the selective pressures acting on them.

Along with a set of sequences, stochastic models of molecular sequence evolution require three elements:

- A phylogeny that indicates the evolutionary relationships between the sequences.
- A multiple sequence alignment that indicates which characters in the sequences are related to each other.
- A set of transition probabilities that give the probability of one sequence mutating into another over a given period of time.

Using these three elements it is possible to compute the likelihood of observing the sequences, assuming that they evolved from a common ancestor. (Of course, this likelihood is conditioned on the three elements above). Calculating the likelihood alone does not elucidate any of the evolutionary relationships or parameters. To aid in our understanding of the evolutionary process one or more of these three elements are parameterized. By finding the parameters that maximize the likelihood insights may be gained on the relationships between the sequences and the forces that play a part in their evolution. The simultaneous optimization of all three elements is computationally difficult. Instead the sequence alignment is usually fixed *a priori*. A parameterization of the phylogeny and transition probabilities is then optimized.

Depending on the complexity of the model some of the parameters may be fixed. For instance, many implementations fix the topology and only optimize the branch lengths in the phylogeny. What exactly constitutes a good model depends on the application. The model should be as simple as possible, but not an oversimplification. Full reality has infinite dimension, which means that it is highly unlikely that the model used will ever be the true underlying model [Burnham and Anderson 2002]. It should be stressed that the purpose of a model is to determine which inferences are supported by the data and not to reveal the truth and that the likelihood should never be confused with the probability of a model being true [Felsenstein 1981]. Furthermore, one should be aware of the fact that more complicated and parameter-rich models are more likely to overfit. Overfitting happens when the model parameters are adjusted to the noise in the data and not to the underlying patterns.

The set of transition probabilities are arguably the most important part to a model of evolution. In this chapter models for the transition probabilities are discussed to study the evolution of individual codons within an alignment. In most cases the model has to be optimized numerically, because it is impossible to find an analytical solution. Strategies for optimizing the model parameters are not discussed here. Finally, it should be noted that any model can be misled by a number of factors, such as a lack of data or unknown processes that are not being modeled.

3.1 Calculating the likelihood

Before the optimization process can begin the likelihood function needs to be computed. The likelihood function can then be maximized by adjusting the model parameters through any numerical optimization algorithm. Let \mathcal{D} be a multiple sequence alignment and \mathcal{T} the topology of the phylogeny. Define θ to be the parameter vector of the model describing the branch lengths of the phylogeny and the parameters of the transition probabilities. The likelihood of observing the data is:

$$\mathcal{L}(\theta|\mathcal{D}, \mathcal{T}) = \Pr(\mathcal{D}|\theta, \mathcal{T})$$

3.1.1 Independent sites

To simplify matters it is usually assumed that the sequence alignment consists of n sites, $\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_n$, with each site evolving independently, which simplifies the likelihood function to:

$$\mathcal{L}(\theta|\mathcal{D}, \mathcal{T}) = \prod_{i=1}^n \Pr(\mathcal{D}_i|\theta, \mathcal{T})$$

Hence, the likelihood for every site can be calculated independently. For computational reasons it is often easier to calculate and optimize the log-likelihood. Since a logarithm is a monotonically increasing function, a maximum of the log-likelihood corresponds to a maximum of the likelihood. The log-likelihood is given by:

$$\ell(\theta|\mathcal{D}, \mathcal{T}) = \sum_{i=1}^n \log \Pr(\mathcal{D}_i|\theta, \mathcal{T}) \quad (3.1)$$

In most models that have been implemented sites were chosen to be nucleotides, amino acids or codons. The motivations behind these choices are discussed in section 3.2. For now it is assumed without loss of generality that each site corresponds to a codon in the alignment.

3.1.2 Markovian evolution

It is assumed that the evolution of a sequence depends only on its current state, which leads to a first order Markovian model for the transition probabilities. Although not completely accurate Markovian models give a compromise between realism and tractability [Felsenstein 1981]. The entries of the transition matrix needs to be specified before a model can be used. Each entry of the transition matrix gives the probability of one codon mutating into another over some period of time. It is usually assumed that the entries of the transition matrix are time independent, which leads to a homogeneous Markov process:

$$P_{ij}(t) = \Pr(\mathcal{D}_i(t_0 + t) = n | \mathcal{D}_j(t_0) = m) \quad (3.2)$$

where the values of $\mathbf{P}(t)$ do not depend on the starting time t_0 at all. Instead of specifying the entries of the transition matrix it makes more sense to define the model in term of its rate matrix, \mathbf{Q} , which contains the instantaneous substitution rates between any two codons. The relationship between the transition and rate matrices is as follows:

$$dP_{ij} = Q_{ij}dt \quad (3.3)$$

The above differ

$$\mathbf{P}(t) = e^{\mathbf{Q}t} \quad (3.4)$$

assuming that $\mathbf{P}(0) = \mathbf{I}$. Hence, the transition matrix for any period of time may be obtained from the rate matrix. However, the transition probabilities only depend on a product of the instantaneous substitution rates and the time of evolution. Hence, it is impossible to uniquely estimate both the rates (values of \mathbf{Q}) and the time without any additional information [Felsenstein 1981]. The confounding of the rates and times may be illustrated by a simple example. Consider halving all the entries of \mathbf{Q} and

doubling the time. This would not change the values in \mathbf{P} at all! The confounding of rates and times is usually sidestepped by arbitrarily setting one of the rates to 1 and then estimating all other rates relative to it.

In order to ensure that \mathbf{P} is a stochastic matrix the entries of \mathbf{Q} are under the constraint that each row has to sum to zero. Hence, the diagonal entries, are usually not specified, but taken as the negative sum of all the other entries in the row. This ensures that the total flow is conserved, since the sum corresponds to the rate at which a codon leaves its current state.

If the frequency of codon i is $\pi_i(t_0)$ at time t_0 , then it is clear that under the model described here, the frequency of codon i will be $\mathbf{P}(t)\pi(t_0)$ at time $t_0 + t$. If the process allows any state to change into any other state it is irreducible and has a stationary distribution [Anisimova and Kosiol 2009]. The stationary distribution of \mathbf{P} is the distribution of codon frequencies such that $\pi_i = \mathbf{P}\pi_i$. Hence, once the stationary distribution has been reached, the codon frequencies will never diverge from it again. The codon frequencies in the stationary distribution π_i , are referred to as the equilibrium frequencies.

The entries of \mathbf{Q} are usually specified to ensure reversibility, which makes it easier to calculate the likelihood on a tree. Reversibility means that the evolutionary process is the same, regardless of whether it is moved forward or backward in time [Felsenstein 1981]. A necessary condition for reversibility is that:

$$\pi_i Q_{ij} = \pi_j Q_{ji}$$

However, it is sufficient to show that \mathbf{Q} can be decomposed into the product of a symmetric and a diagonal matrix:

$$\mathbf{Q} = \mathbf{R}\mathbf{\Pi}$$

where the off-diagonal entries of \mathbf{R} contain the codon exchangeabilities and $\mathbf{\Pi}$ is the diagonal matrix containing the equilibrium frequencies.

Finally, the entries of \mathbf{Q} need to be scaled to ensure a mean substitution rate of 1 at equilibrium [Anisimova and Kosiol 2009]. This is done by finding the scaling factor μ such that:

$$-\sum_i \mu \pi_i Q_{ii} = 1$$

The result of the scaling is that the length of the branches are measured in the expected number of substitutions per site.

3.1.3 Independent branches

Using the definitions from the section above it is possible to calculate the probability of a codon evolving into another along any of the branches in the tree. Calculating the likelihood of observing the sequences on the whole tree involves some more work. As sites are assumed to be independent the likelihood can be calculated for each site separately and then combined into the total likelihood with equation 3.1. In order to calculate the likelihood of observing the different sequences on the tree, the assumption is made that after divergence, evolution proceeds independently along the two lineages [Felsenstein 1981]. For any branch in a tree, given the states at the ends of the branch, the likelihood on the tree is equal to the likelihood of the state change across the branch, multiplied by the likelihoods on the two subtrees anchored at the ends of the branch. Assuming that the states of all the internal nodes to the tree in figure 3.1 are known, the likelihood of site i can be written as:

$$\mathcal{L}(\theta|\mathcal{D}_i, T) = \pi_A P_{AX}(d_1) P_{XB}(d_2) P_{XY}(d_6) P_{YE}(d_5) P_{YZ}(d_7) P_{ZC}(d_3) P_{ZD}(d_4)$$

The likelihood computation is started arbitrarily from any of the nodes in the tree. By multiplying together the frequency of the codon at the starting node and the probability of the changes observed across all the branches in the tree the total probability is obtained. The reversibility property ensures that the computation can be started from any node [Felsenstein 1981]. In practice the states of the internal nodes are not known, hence the likelihood has to be summed across all possible combinations:

$$\mathcal{L}(\theta|\mathcal{D}_i, T) = \sum_X \sum_Y \sum_Z \pi_A P_{AX}(d_1) P_{XB}(d_2) P_{XY}(d_6) P_{YE}(d_5) P_{YZ}(d_7) P_{ZC}(d_3) P_{ZD}(d_4)$$

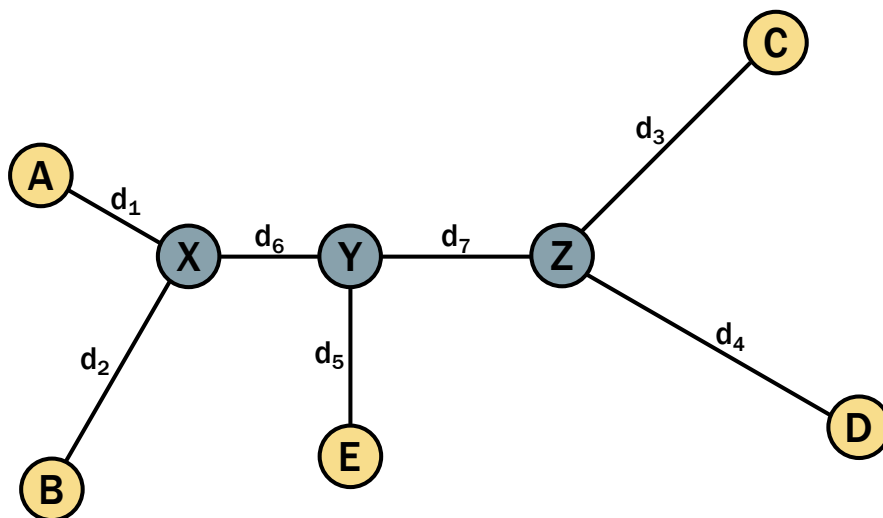


Figure 3.1: An example phylogeny.

This is a very costly computation, but it can be optimized significantly by Felsenstein’s pruning algorithm, which basically consists of moving all the summations inward as far as possible to remove unnecessary terms [Felsenstein 1981].

3.2 Codon models

When modeling evolution on coding sequences it does not make sense to use a nucleotide model, since evolutionary pressures act on the amino acids and codons. A codon model gives a higher resolution than an amino acid model and makes it possible to model the effect of synonymous (or silent) substitutions that do not affect the amino acid sequence of a gene. Such substitutions are thought to play an important role in the evolutionary process (see chapter 2), although the effect of nonsynonymous substitutions will always be greater. Moreover, it has been shown that even when comparing distant organisms, where synonymous substitutions may be saturated, a codon model still gives a better fit than an amino acid model [Anisimova and Kosiol 2009]. Furthermore, codon models have the ability to distinguish between nonsynonymous and synonymous substitutions, which can be used to measure the selective pressure on a site [Anisimova and Kosiol 2009]. Diversifying (or positive) selection acts on sites where changes to the amino acid sequence are advantageous and is characterized by an overabundance of nonsynonymous substitutions. On the other hand, a site that is under purifying (or negative) selection shows a deficit of nonsynonymous substitutions, which means that the amino acid sequence is preserved.

An obvious drawback of codon models is the extra state space that needs to be modeled. Where the rate matrices for nucleotide and amino acid models only have to model transitions between respectively 4 and 20 possible states, a codon model needs to handle transitions between 64 different states. This is usually simplified a little, by assuming that any mutations to and from stop codons are extremely deleterious and can safely be disregarded [Goldman and Yang 1994]. Nevertheless, the rate matrix still has a size of 61×61 , which is considerably larger than the rate matrices for nucleotide and amino acid models. As a result codon models are much more challenging computationally and it usually takes much longer to optimize them.

3.2.1 Rate matrices

The first two codon models were introduced simultaneously by Goldman and Yang [1994] and Muse and Gaut [1994]. From the assumption that mutations occur independently at different codon positions it follows that the probability of substitutions between codons that differ in more than one nucleotide are an order of magnitude smaller than substitutions between codons that differ in only one nucleotide. Hence, the substitution rate between any two codons differing in more than one nucleotide is set to

zero in both models. This still allows for substitutions between any two codons, just not in one time step [Goldman and Yang 1994].

Additional parameters may be introduced to account for the nucleotide change in the codon substitution. In the Goldman-Yang model all transitions are multiplied by the parameter κ , the transition-transversion ratio. Although nucleotide biases are not modeled in the original Muse-Gaut model later implementations can be coupled with any of a number of different nucleotide models, such as HKY85 or the general time reversible model. The most important parameters modeling the codon changes are for the nonsynonymous and synonymous substitution rates, d_N and d_S . Hence, synonymous changes occur at a different rate to nonsynonymous changes.

Originally, the Goldman-Yang model also incorporated differences between amino acids in nonsynonymous substitutions by using Grantham distances [Goldman and Yang 1994]. However, these distances were not found to be very informative and were subsequently dropped from the model [Yang and Nielsen 1998]. Although it has been shown that there is a large bias in the exchangeabilities between amino acids based on their physico-chemical properties these differences are almost never accounted for in codon models. Nevertheless, it has been found that codon models still fit data better than amino acid models that account for these differences [Anisimova and Kosiol 2009].

The only real difference between the two models lies in their treatment of the equilibrium frequencies. In the Goldman-Yang model a substitution is proportional to the frequency of the target codon, π_j , whereas in the Muse-Gaut model it is proportional to the frequency of the target nucleotide. That is, if the two codons differ at nucleotide k , then the substitution is proportional to $\pi_{j_k}^{nu}$. The two models can be written as:

$$Q_{ij}^{\text{GY}} = \begin{cases} 0 & \text{if } i \text{ and } j \text{ differ in more than one nucleotide} \\ \alpha r^{nu}(i_k, j_k) \pi_j & \text{for a synonymous substitution} \\ \beta r^{nu}(i_k, j_k) \pi_j & \text{for a nonsynonymous substitution} \end{cases} \quad (3.5)$$

$$Q_{ij}^{\text{MG}} = \begin{cases} 0 & \text{if } i \text{ and } j \text{ differ in more than one nucleotide} \\ \alpha r^{nu}(i_k, j_k) \pi_{j_k}^{nu} & \text{for a synonymous substitution} \\ \beta r^{nu}(i_k, j_k) \pi_{j_k}^{nu} & \text{for a nonsynonymous substitution} \end{cases} \quad (3.6)$$

where $r^{nu}(a, b)$ specifies the substitution bias between two nucleotides, α the synonymous substitution rate and β the nonsynonymous substitution rate.

It is not clear which of the two models gives a better representation of the data. Although at first it seems clear that codon frequencies should be used as evolution on codons is being modeled, some researchers have pointed out that the Goldman-Yang model weights a single nucleotide substitution with the frequency of the whole codon, which could lead to undesired effects [Yap *et al.* 2010, Rodrigue *et al.* 2008]. It is argued that this could result in context effects when there are none, which could in turn lead to estimates predicting the presence of selection when the process is neutral. Yap *et al.* [2010] further argue that the Muse-Gaut model ignores context effects from neighbouring nucleotides within a codon. Another consequence of the Muse-Gaut model is that codon frequencies are multiplicative:

$$\pi_i = \pi_{i_1}^{nu} \pi_{i_2}^{nu} \pi_{i_3}^{nu}$$

It is argued that this will hardly ever be the true scenario [Yap *et al.* 2010].

Based on the above deficiencies, Yap *et al.* [2010], defined a new codon model, where substitutions are proportional to the frequency of the target nucleotide conditioned on the frequencies of the other nucleotides in the target codon. Hence, in their model the frequency weighting for a substitution between two codons is:

$$\pi^{\text{YAP}}(i, j) = \begin{cases} \pi_{j_1|j_2, j_3} & i_1 \neq j_1, i_2 = j_2, i_3 = j_3 \\ \pi_{j_2|j_1, j_3} & i_1 = j_1, i_2 \neq j_2, i_3 = j_3 \\ \pi_{j_3|j_1, j_2} & i_1 = j_1, i_2 = j_2, i_3 \neq j_3 \end{cases} \quad (3.7)$$

This model is related to the Muse-Gaut model in the sense that the two models are equal when codon frequencies are multiplicative. Further, it is also possible to write the Goldman-Yang model in terms of the Yap or the Muse-Gaut model, by dividing the codon exchangeabilities with extra frequency parameters [Yap *et al.* 2010, Lindsay *et al.* 2008]. The authors argue that this is evidence of

the Goldman-Yang model introducing a bias and confounding rate and frequency parameters. They further go on to show that in both simulations and real data both the Goldman-Yang and Muse-Gaut models detect selection when applied to neutrally evolving sequences, while the Yap model makes the correct inferences. However, in their simulations they used data simulated according to their model. Hence, their test merely proves that their implementation is correct, and not that their model is superior. Furthermore, although the real data they used (primate intron sequences) are thought to be evolving neutrally, there may be other biases at work as well, which casts some doubts on their conclusions.

3.2.2 Parameter estimation

Due to the confounding of rates and times one of the rates is always set equal to one. In practice, only the ratio of nonsynonymous and synonymous substitutions, ω is estimated. This can be used to detect selection. A sequence is under positive selection if $\omega > 1$ and negative selection if $\omega < 1$. It is further also necessary to set one of the rates in the nucleotide model to one.

The Goldman-Yang model has 60 estimable frequency parameters, while the Muse-Gaut model only has 3. (Since the frequencies sum to one it is not necessary to estimate all the parameters). These formulations are respectively referred to as F61 and F1 \times 4. A refinement to the Muse-Gaut model uses the position specific frequencies of nucleotides in the target codon. That is, if the substitution is between CGT and CTT, then the frequency of T in second codon positions will be used, instead of the frequency of T in all codon positions. This results in 9 estimable frequency parameters, referred to as F3 \times 4. The Yap model may be used with either F1 \times 4 or F3 \times 4.

The Goldman-Yang model usually achieves a better likelihood and the frequencies obtained by it are usually much closer to the observed codon frequencies than those obtained by the other models [Yap *et al.* 2010]. However, these advantages come at the price of many extra parameters. In most cases the alignment does not contain enough information to directly estimate the frequency parameters, hence the empirical frequencies are often used [Kosakovsky Pond *et al.* 2010a]. The empirical frequencies are found by counting the observed frequencies in the data, according to F61, F1 \times 4 or F3 \times 4. For instance, in the case of F3 \times 4 the frequency for a codon is given by:

$$\pi_j = \frac{\pi_{j_1}^{nu_1} \pi_{j_2}^{nu_2} \pi_{j_3}^{nu_3}}{1 - \sum \pi_{\text{stop}}} \quad (3.8)$$

where the correction in the denominator accounts for the fact that stop codons are not modeled. Even for the Goldman-Yang model empirical frequencies are often calculated by F3 \times 4 since there usually are not enough data to find reliable empirical estimates of the codon frequency. Note that doing so removes the likelihood advantage that the Goldman-Yang model had over the other models. A recent modification to F3 \times 4, called CF3 \times 4 was introduced to correct for a bias caused in disregarding the stop codons by solving a system of non-linear equations [Kosakovsky Pond *et al.* 2010a]. Finally, when the empirical frequencies are used the frequency parameters do not count toward the free parameters of the model, as only parameters that are uniquely estimable from the data should be counted toward the degrees of freedom [Burnham and Anderson 2002].

3.3 Models with rate variation

All the models described until now have been globally homogeneous. That is, the same rate matrix is used on all sites and branches. Henceforth, such models are referred to as constant rates models. A constant rates model, such as the ones described in the previous section, estimates the mean ω rate across all sites and lineages. In most sites an excess of nonsynonymous substitutions would be deleterious. Therefore, it is thought that the majority of genes evolve under purifying selection and that positive selection affects only a few sites and lineages [Anisimova and Kosiol 2009]. In these cases a constant rates model would not have enough power to detect the presence of positive selection [Yang *et al.* 2000]. This led to the introduction of locally homogeneous models, or models with rate variation. By comparing the fit of a model of rate variation to a constant rates model it is possible to test for the presence of selection at specific codon sites [Nielsen and Yang 1998].

Rate variation may be introduced in two ways. In the first formulation the rate matrix is allowed to vary over time. In this case each branch of the phylogeny may have a different rate matrix, as in [Yang and Nielsen 1998]. These models are not discussed here. Instead this research focusses on the second formulation of rate variation, where the rates are allowed to vary from site-to-site along the alignment. It is also possible to have rate variation across both sites and lineages [Nielsen and Yang 2003], however such models are very rich in parameters, and easily overfit.

3.3.1 Formulation of site-to-site variation

It is likely that there are different rates for every site in the alignment. Suppose for now that the only rate that varies is ω . To model rate variation, ω is drawn from a continuous distribution, $f(\omega)$. To calculate the likelihood at site i , it now becomes necessary to integrate over all possible values for ω :

$$\Pr(\mathcal{D}_i) = \int_0^\infty \Pr(\mathcal{D}_i|\omega)f(\omega)d\omega$$

However, Yang [1993] found that using a continuous distribution makes the computations intractable. Instead, a discrete distribution is used, so that a model with k discrete rate classes has k discrete values of ω . More generally, a model with k discrete rate classes has k different rate matrices, $\mathbf{Q}_1, \mathbf{Q}_2, \dots, \mathbf{Q}_k$. Instead of assigning a site to a rate matrix *a priori*, each rate matrix is assigned a weight and the likelihood calculation for a site becomes the weighted average of the likelihood for a site given each of the rate matrices:

$$\Pr(\mathcal{D}_i) = \sum_{j=1}^k \Pr(\mathcal{D}_i|\mathbf{Q}_j) \Pr(\mathbf{Q}_j) \quad (3.9)$$

It is still necessary to specify a distribution from which the rates are drawn. A general discrete distribution (GDD) or a discretized continuous distribution may be used. In GDD models no assumptions are made about the distribution aside from the number of rate classes. Both the rates in the different classes and the weights of the different rate classes are estimated, adding an extra $2k - 1$ free parameters to the model if only one rate is assumed to vary. (Only $k - 1$ weights need to be estimated since the weights sum to 1).

Among continuous distributions a Γ distribution is the most widely used, perhaps because the original formulation used one [Yang 1994]. However, several other distributions and mixtures of distributions have been examined and it has been found that other distributions, in particular a β distribution, give a better fit than a Γ distribution [Yang *et al.* 2000, Kosakovsky Pond and Frost 2005a].

Regardless of the distribution used, the discretization procedure stays the same. Assuming that there is only one variable rate, for instance ω , the distribution is discretized by finding a sequence of numbers, $a_1 = 0, a_2, \dots, a_k = \infty$ such that:

$$\Pr(\omega_i \in [a_{j-1}, a_j]) = \int_{a_{j-1}}^{a_j} f(\omega)d\omega = p_{\omega_i} \quad (3.10)$$

where the value of p_{ω_i} is usually decided *a priori* and denotes the weight of rate class i . The rate of ω used for a rate class is taken as either the mean or the median rate within the interval of its rate class. When using a discretized continuous distribution the assumption is usually made that all rate classes have the same weight. Hence, only the parameters of the distribution need to be estimated during optimization. This makes using a discretized continuous distribution attractive since it does not add many free parameters to a model. General discrete distributions are more difficult to fit and usually do not converge for more than 4 rate classes [Kosakovsky Pond and Frost 2005a].

3.3.2 Models of site-to-site variation

In codon models the choice is usually made to only allow variation in the selective pressures. In the first implementation of codon models with site heterogeneity Nielsen and Yang [1998] defined two models using a GDD. The first model is used to test for the neutral theory and has two categories, one where ω is restricted to be 1 and one where ω is allowed to be negative. The second model tests for

positive selection and has an extra category where ω is allowed to be greater than one. Later, more general models that allow ω to vary freely along different distributions, and models that assign some sites to an invariable class were introduced (see for instance [Yang *et al.* 2000, Kosakovsky Pond and Frost 2005a]).

In all of these implementations ω is the only parameter that is allowed to vary. Estimating ω as the d_N/d_S ratio is effectively the same as setting the synonymous rate to 1 [Kosakovsky Pond and Muse 2005]. Consequently, even when ω is allowed to vary, it is actually only d_N that varies, while d_S remains fixed. Traditionally it was assumed that the synonymous substitution rate is neutral and hence does not vary [Chamary *et al.* 2006]. However, evidence has been mounting that there is also selection on synonymous sites and that this leads to a variation in the synonymous rate (see section 2.2). This led to the introduction of the dual model by Kosakovsky Pond and Muse [2005].

In the dual model both α and β , as defined in equation 3.6, are allowed to vary. As a result both nonsynonymous and synonymous rate variations are modeled. The dual model further makes the assumption that α and β are drawn from independent distributions. To prevent the confounding of rates and times the mean synonymous rate is arbitrarily set to 1. Effectively this means that if a model has n nonsynonymous and m synonymous rate classes, it will have a total of mn rate classes, and the site specific likelihood calculation becomes:

$$\Pr(\mathcal{D}_i) = \sum_{j=1}^m \sum_{k=1}^n \Pr(\mathcal{D}_i | \alpha_i = \alpha_j, \beta_i = \beta_k) \Pr(\alpha_i = \alpha_j, \beta_i = \beta_k)$$

where:

$$\Pr(\alpha_i = \alpha_j, \beta_i = \beta_k) = \Pr(\alpha_i = \alpha_j) \Pr(\beta_i = \beta_k) \quad (3.11)$$

since the distributions are independent. Furthermore, when discretizing continuous distributions, the distributions for α and β can be discretized independently by equation 3.10, resulting in a rectangular grid partitioning the combined bivariate distribution.

Kosakovsky Pond *et al.* [2010b] states that assuming that the nonsynonymous and synonymous rates are uncorrelated is very inefficient, since the number of rate classes is equal to the product of the individual rate classes. The number of likelihood evaluations that need to be performed is linear with respect to the number of rate classes which means that these models do not scale well. Furthermore, the assumption of independence means that the discrete rate values are constrained to fall on a rectangular grid on the underlying bivariate distribution, with many of the rate classes being assigned very small weights. To remedy this situation they propose the general bivariate discrete distribution (GBDD) model, which draws α and β from a general bivariate discrete distribution. They claim that a model that allows the rates to covary will perform as well as or better than a model that imposes an artificial constraint of independence. The calculation of the site-specific likelihood remains unchanged from the dual model, with the added advantage of fewer rate classes being able to model the same amount of variation [Kosakovsky Pond *et al.* 2010b]. (However, equation 3.11 no longer holds). A drawback of this model is that the added freedom it has in choosing rates makes it much more unstable.

3.4 Models that account for codon usage biases

The observations described in the previous chapter indicate that codon biases may be under selection. In [Nielsen *et al.* 2007] a selection coefficient is introduced to model the effect of mutations from preferred to unpreferred codons and vice-versa. Instead of dividing codons into two classes the FMutSel model of Yang and Nielsen [2008] goes one step further and adds a fitness parameter, f_i , for every codon i . The selection coefficient for a change between codons i and j is then:

$$s_{ij} = f_j - f_i$$

If the effective population size is N_e , then the probability of fixation of the mutation is:

$$\frac{2s_{ij}}{1 - e^{-2N_e s_{ij}}}$$

By introducing the rescalings $F_i = 2N_e f_i$ and $S_{ij} = 2N_e s_{ij} = F_j - F_i$, it can be shown that the instantaneous substitution rate between codon i and j is:

$$Q_{ij} \frac{S_{ij}}{1 - e^{-S_{ij}}} = Q_{ij} h(S_{ij})$$

For neutral mutations $S_{ij} = 0$. Similarly $S_{ij} < 0$ for deleterious mutations and $S_{ij} > 0$ for advantageous mutations. This formulation adds 61 extra parameters to the model. However, only 60 of these parameters are estimable, since the model only depends on the differences, $F_j - F_i$. To remedy this, one of the fitness parameters are arbitrarily set to 0. It should also be noted that without any information on the effective population size only the scaled fitnesses can be estimated. This model is still reversible, although the proof of reversibility is quite involved [Yang and Nielsen 2008]. The model by [Nielsen *et al.* 2007] is exactly the same as FMutSel, except for the fact that the codons are divided into preferred and unpreferred classes. Since there are only two classes, only one extra parameter can be estimated, and hence this is left as the selection coefficient and not interpreted as the fitness of preferred or unpreferred codons.

FMutSel is difficult to fit because of the large amount of free parameters. On the other hand, the model of Nielsen *et al.* [2007] only allows for the selection of preferred codons. This may be an unreliable assumption as there is evidence of selection for unpreferred codons in some cases (see section 2.2). In the model of Zhou *et al.* [2010] codons are again divided into preferred and unpreferred codons. However, the model assumption is not that preferred codons have a higher fitness, but that selection favours the preservation of the codon status. Hence, synonymous substitutions are divided into two classes. Conservative synonymous substitutions occur between two preferred or two unpreferred codons. Nonconservative synonymous substitutions change the status of the codon from preferred to unpreferred or vice versa. This leads to the introduction of two new evolutionary rates, d_{S_C} and d_{S_N} , respectively the conservative and nonconservative synonymous rates. Due to the confounding of rates and times only the ratio $\psi = d_{S_N}/d_{S_C}$ can be estimated. This model can be added on to any of the previously defined codon models, for instance the Goldman-Yang model:

$$Q_{ij} = \begin{cases} 0 & \text{if } i \text{ and } j \text{ differ in more than one nucleotide} \\ r^{nu}(i_k, j_k)\pi_j & \text{for a conservative synonymous substitution} \\ \psi r^{nu}(i_k, j_k)\pi_j & \text{for a nonconservative synonymous substitution} \\ \omega r^{nu}(i_k, j_k)\pi_j & \text{for a nonsynonymous substitution} \end{cases}$$

The ψ parameter can be used to detect purifying ($\psi < 1$) or diversifying selection ($\psi > 1$) on synonymous substitutions. Zhou *et al.* [2010] further argue that the conservative synonymous substitution rate may be used as a more accurate estimate of the neutral substitution rate. A drawback of the models of Zhou *et al.* [2010] and Nielsen *et al.* [2007] is that the preferred and unpreferred codons need to be specified first.

3.5 Discussion

The type of models described in this chapter are sometimes referred to as random effects likelihood (REL) models, since they do not assign sites to rate classes *a priori*. Fixed effects likelihood (FEL) models partition sites based on previous knowledge [Bao *et al.* 2007]. Empirical codon models are a different breed of model where the transition matrix is estimated from a large dataset and then simply applied to new datasets [Schneider *et al.* 2005, Kosiol *et al.* 2007]. These models usually give a better fit, since they can capture some subtleties and effects not modeled explicitly by a parametric model. On the other hand, empirical models cannot capture the variation in the substitution rate between different genes, since no parameters are optimized [Muse and Gaut 1994]. Before the advent of stochastic codon models, counting methods, as used in [Suzuki and Gojobori 1999, Gaut *et al.* 1996], provided the only way to estimate the nonsynonymous and synonymous rates computationally. Although these methods are fast, they have several drawbacks. They cannot model multiple substitutions and require a large amount of data for accurate inferences. Furthermore, they can only be used in a pairwise fashion [Yang and Nielsen 1998]. Kosakovsky Pond and Frost [2005b] found similar results with REL, FEL and counting methods, however this depends on the focus of the study and the amount and quality of the data.

Chapter 4

Implementation and validation of stochastic codon models

At present there are no commercial software packages tailored to the simulation and maximum-likelihood estimation under Markovian codon models. The tools that are available are often open source and only supported by the scientific community. One of the biggest drawbacks to this situation is that none of the available tools implement all of the models discussed in the previous chapter. Furthermore, the available tools often have a steep learning curve, sparse documentation and may be unstable. PAML and HyPhy have emerged as the most popular tools for investigating codon models.

PAML¹ is a suite of programs for the phylogenetic analyses of DNA and amino acid sequences [Yang 2007; 1997]. PAML includes a multitude of nucleotide, amino acid and codon models of evolution. Codon models of evolution are implemented in the `codeml` program within the PAML package. HyPhy² is a very flexible package for studying all aspects of molecular evolution [Kosakovsky Pond *et al.* 2005]. The HyPhy batch language is a high-level programming language that may be used to specify custom models or analyses in HyPhy. A large number of batch files are available that implement all of the standard analyses researchers are likely to carry out. A GUI version of HyPhy is also available that makes it easy for beginners to get started. Unlike PAML, HyPhy also includes models that allow the synonymous rate to vary.

PyCogent³ is a similar package implemented as a set of Python libraries [Knight *et al.* 2007]. Unlike PAML and HyPhy, PyCogent also implements the Yap frequency model. The fact that it is part of a comparative genomics package makes it easy to use for pre- and postprocessing of data, further analyses and producing formatted output. Additionally, it is also easy to extend. On the other hand, since PyCogent is part of a Python package and not a standalone tool, it is significantly more difficult to get started with it than any of the other tools.

One thing all these tools have in common is that they are mainly used to fit a model of evolution to a given phylogenetic tree. Due to the complications of exploring the space of possible topologies these packages are not very good at phylogenetic reconstruction. Another package, MrBayes [Ronquist and Huelsenbeck 2003], can perform phylogenetic reconstruction using codon models. MrBayes infers phylogenies according to a Bayesian framework using Monte Carlo simulations, whereas all the other packages described above use maximum likelihood. The fundamental difference between the methodologies is that Bayesian methods do not condition the likelihood on a single tree topology, but instead weights every tree according to its posterior probability.

In Zanetti [2010], CodonPhyML⁴, an extension of PhyML to codon models, is described. (A manuscript is in preparation and will be submitted for publication in the nearby future, [Zanetti *et al.* 2011]). PhyML⁵ is a widely used phylogenetic inference tool [Guindon *et al.* 2010, Guindon and Gascuel 2003]. It is known for being one of the fastest phylogenetic inference tools available while still providing high levels of accuracy, making it ideal for use on large datasets. PhyML implements a large

¹PAML is available at <http://abacus.gene.ucl.ac.uk/software/paml.html>

²HyPhy is available at <http://www.datam0nk3y.org/hyphy/doku.php?id=start>

³PyCogent is available at <http://pycogent.sourceforge.net/>

⁴CodonPhyML is available at <http://sourceforge.net/projects/codonphyml/>

⁵PhyML is available at <http://www.atgc-montpellier.fr/phyml/>

number of nucleotide and amino acid models of evolution. After an initial tree is computed, using a fast distance-based algorithm or parsimony, optimal parameters of the selected model are found by maximum likelihood. (Alternatively, an initial tree may also be provided). Branch length and rate parameters are optimized numerically and the topology can be optimized through heuristic searches, by nearest neighbour interchanges (NNI), subtree pruning and regrafting (SPR) or a combination of the two. After the optimization is completed branch supports may be computed by an approximate likelihood ratio test (aLRT) or a non-parametric SH correction for the aLRT (SH-aLRT) [Guindon *et al.* 2010]. Branch supports may also be computed by bootstrap. (See [Anisimova *et al.* 2011] for a comparison of the performance of these methods).

The fact that none of the available tools for fitting codon models can be used to simultaneously infer a phylogeny, coupled with the popularity and speed of PhyML, motivated its extension to codon models. Prior to the start of this project CodonPhyML implemented three types of rate models for codon evolution, the Constant Rates model and models that allow the nonsynonymous rate to vary, either along a general discrete distribution (Nonsynonymous GDD) or along a Γ distribution (Nonsynonymous Γ). (These models are respectively referred to as M0, M3 and M5 by Yang *et al.* [2000], and this notation is also used in [Zanetti 2010]). Aside from parametric codon models, CodonPhyML also implements several empirical and semi-empirical codon models. The Goldman-Yang, Muse-Gaut or Yap frequency models can be coupled with any of the codon models. Lastly, CodonPhyML still implements all the nucleotide and amino acid models implemented in PhyML.

In PhyML, branch lengths and rate parameters are optimized one by one using Brent’s algorithm [Zanetti 2010]. This was found to be inefficient for codon models with more than one rate category and hence was changed to the BFGS algorithm used in codeml [Zanetti 2010]. In addition to the above codon models, the extension of PhyML to CodonPhyML also involved the addition of several other features. Three different methods to compute the matrix exponential (required to calculate the transition matrix) are available in CodonPhyML [Zanetti 2010]. Besides the standard eigen-decomposition two fast approximate methods are also available, in the form of a Padé approximation and a Taylor series approximation. The running time of CodonPhyML was improved by using a Taylor series approximation for the rate matrix while optimizing rate parameters, and by using routines from BLAS/LAPACK to speed up computations where applicable. OpenMP was also used to parallelize several loops. One final addition to CodonPhyML is the possibility to compute the initial tree using the empirical codon models of Kosiol *et al.* [2007] and Schneider *et al.* [2005].

4.1 Additions to CodonPhyML

The main contribution made to CodonPhyML was the introduction of models that allow both the synonymous and the nonsynonymous rates to vary. Three such models were added, the Dual model with a general discrete distribution (Dual GDD) or a Γ distribution (Dual Γ) (see [Kosakovsky Pond and Muse 2005]) and the general bivariate discrete distribution (GBDD) model (see [Kosakovsky Pond *et al.* 2010b]). Any of the three available frequency models can be coupled with these models.

The model described by Zhou *et al.* [2010], that allows for selection on the codon usage was also implemented. This model can be combined with any of the other parametric codon models to allow for the effect of selection on conservative or nonconservative synonymous substitutions.

Lastly, a post-processing step was added to CodonPhyML to output site-specific likelihood values, posterior rates and probabilities and rate class assignments.

4.1.1 Implementation of models that allow synonymous rate variation

The major difference between models that allow for rate heterogeneity and the Constant Rates model is that they have more than one rate matrix. All of the rate matrices for parametric models are constructed in CodonPhyML by one of three functions, depending on the frequency model used (Update_Qmat_GYSIMP, Update_Qmat_MGSIMP or Update_Qmat_YAPSIMP). In facilitating synonymous rate variation, the minimum number of changes were made to these functions, with only small additions elsewhere in the code. In the Nonsynonymous GDD and Nonsynonymous Γ models that were already implemented, the only difference between the rate matrices lies in the ω values. For these models d_S is fixed to 1, which means that $\omega = d_N$.

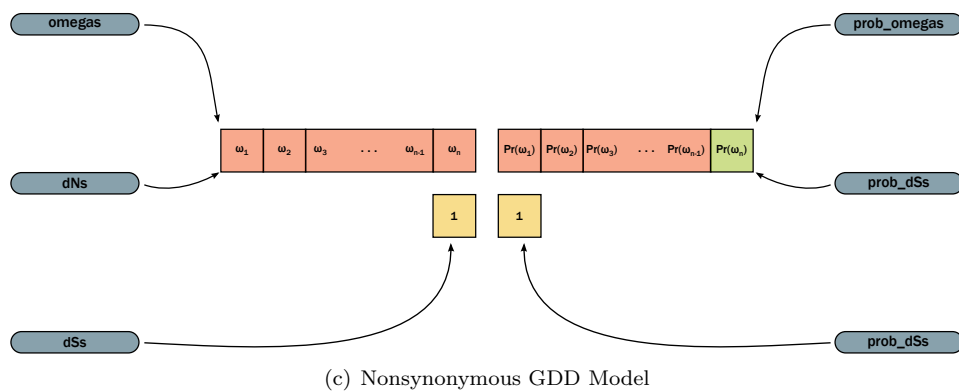
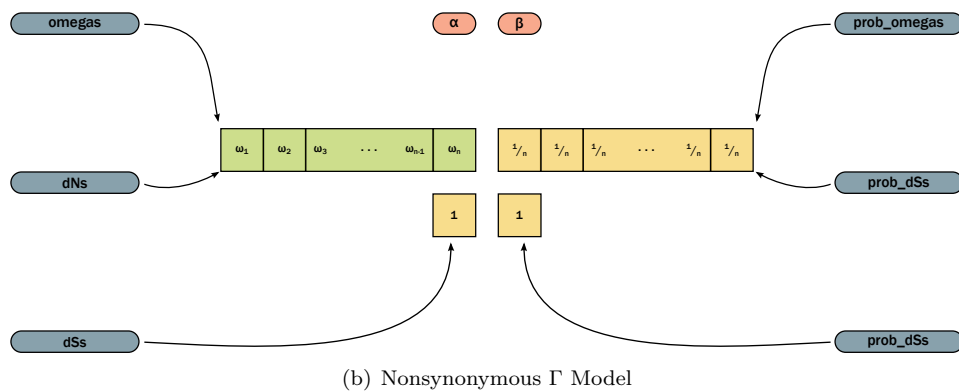
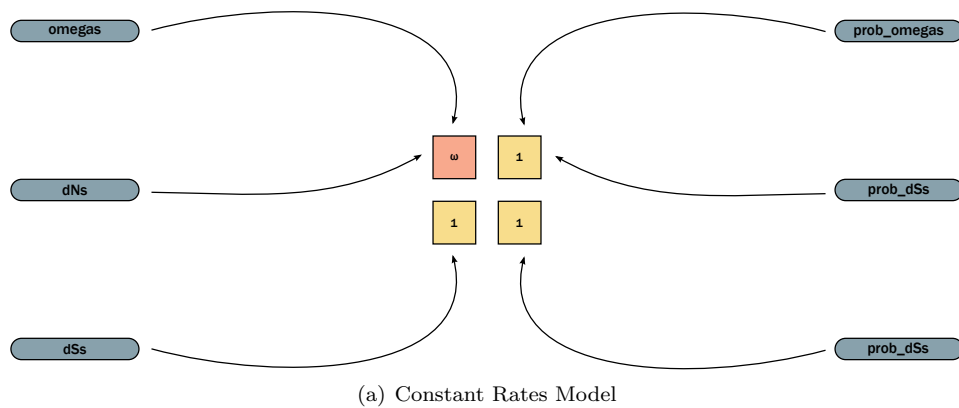
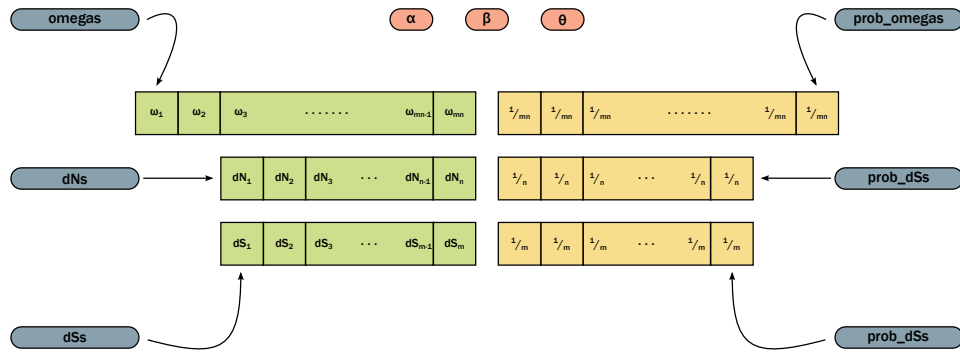
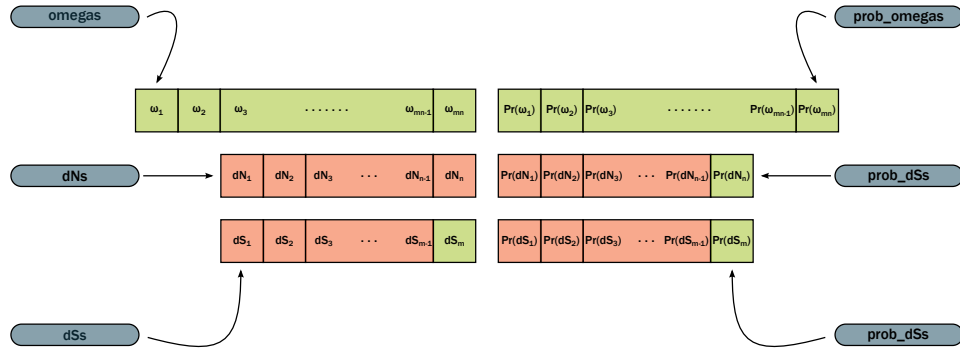


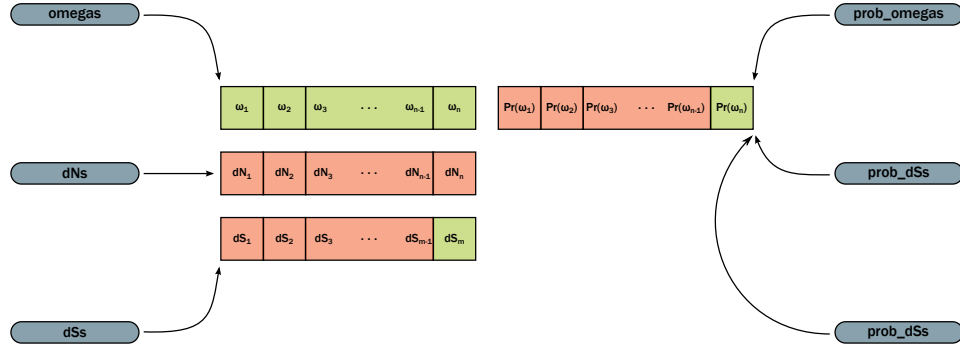
Figure 4.1: Variables stored in memory that are related to the nonsynonymous and synonymous rates of evolution for the three models that were previously implemented in CodonPhyML. Pointers are blue, variables that are optimized with BFGS are red, variables that are derived from other variables are green and variables that are fixed *a priori* are yellow. Note that only the variables in red add toward the free parameters of a model.



(a) Dual Γ Model



(b) Dual GDD Model



(c) GBDD Model

Figure 4.2: Variables stored in memory that are related to the nonsynonymous and synonymous rates of evolution for the three models that were added to CodonPhyML. Pointers are blue, variables that are optimized with BFGS are red, variables that are derived from other variables are green and variables that are fixed *a priori* are yellow. Note that only the variables in red add toward the free parameters of a model and that the dual models may have an unequal amount of nonsynonymous and synonymous rate classes. The θ variable for the Dual Γ model is the single parameter for the Γ distribution used to fit the synonymous rates.

Therefore, allowing for variation in both d_N and d_S required introducing only one new parameter for d_S and renaming ω to d_N in the implementation. (What the variable is called is arbitrary). This formulation allows the construction of rate matrices with different nonsynonymous and synonymous rates. In the implementation, CodonPhyML always stores arrays for the rates and distributions of ω , d_N and d_S in memory. The situation in memory for the different rate models is illustrated in figures 4.1 and 4.2. In the Constant Rates and Nonsynonymous models the same array is used to store ω and d_N rates. In these models only one synonymous rate class is defined with the rate fixed to 1 *a priori*. As can be seen from the figure the additional memory usage is minimal. In the Dual models both d_N and d_S are allowed to vary independently. Thus, if there are n nonsynonymous rates and m synonymous rates, this model requires $n \times m$ rate matrices to be defined, for every combination of classes. The situation for the GBDD model is a combination of the dual and nonsynonymous models. Although the rates vary independently there is only one set of weights. It can be seen from the figures that the three GDD models have far more free parameters than the other models.

In the Dual and GBDD models ω rate ratios are calculated after each round of optimization as:

$$\omega_k = \frac{d_{N_i}}{d_{S_j}}$$

In the Dual models the nonsynonymous and synonymous rates are independent allowing the weights of the ω rates to be calculated as:

$$\Pr(\omega_k) = \Pr(d_{N_i}) \Pr(d_{S_j})$$

Originally, the ω values were used in the construction of the rate matrices, but this was found to be too unstable. Instead fractions are avoided in its construction, as the estimation of fractions is problematic [Kosakovsky Pond and Muse 2005].

In the models that make use of a general discrete distribution (Nonsynonymous GDD, Dual GDD, GBDD), the rates and their respective weights are optimized with BFGS along with the other parameters. In the case of the Nonsynonymous GDD model only the nonsynonymous rates and their weights are optimized. In the Dual GDD and GBDD models the synonymous rates are also optimized. Due to the confounding of rates and times the mean synonymous rate is always fixed to 1 (see section 3.3.2). This is implemented as follows. If there are m synonymous rate classes, only the first $m - 1$ synonymous rates are optimized. After the optimization step, the last synonymous rate is set to 1 and all of the synonymous rates are normalized according to:

$$d_{S_k} = \frac{d_{S_k}}{\sum_{i=1}^m d_{S_i} \Pr(d_{S_i})}$$

In GDD models with n rates, only the weights of $n - 1$ rates need to be optimized, since the weights sum to 1. In the models that make use of Γ distributions only the parameters of the distributions are optimized and not the rates themselves. For the Nonsynonymous Γ model there is only one Γ distribution:

$$f(x, \alpha, \beta) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}$$

that is used to model the nonsynonymous variation. The two parameters, α and β are optimized with BFGS. The Dual Γ model adds a second Γ distribution to model synonymous variation. Due to the confounding of rates and times α and β are set equal in this distribution to force the mean synonymous rate to 1, which leads to only 3 independent rate parameters (see section 3.3.2)⁶. In both cases the rates are obtained after the optimization by discretizing the Γ distribution according to equation 3.10. In all the models using Γ distributions all the rate classes are assigned equal weights.

4.1.2 Models that allow selection on codon usage

In CodonPhyML, a 64×64 matrix is stored in memory that holds the properties of a substitution between any two possible codons. For every pair of codons this matrix stores three properties:

⁶The mean of the Γ distribution equals α / β

- The number of nucleotides that differ between the codons.
- If the difference is a transition or a transversion (only meaningful if the number of differences is equal to one).
- If the substitution is synonymous or nonsynonymous.

This matrix is consulted when the rate matrix is built to determine the right type of substitution for each entry.

The model by Zhou *et al.* [2010] divides codons into two classes, preferred and unpreferred. Synonymous substitutions between the classes are termed nonconservative and multiplied by a rate ratio ψ . (See section 3.4 for more details). This was easily added to CodonPhyML by dividing the synonymous substitutions into two different classes. Thus, a substitution may be nonsynonymous, synonymous conservative or synonymous nonconservative. This extension can be added on to any of the previously defined codon models, which is indicated by adding $+\psi$ to the model description, for example Dual GDD $+\psi$. Furthermore, this formulation is easy to extend to models with more classes of substitutions.

4.1.3 Post-processing

An optional post-processing step was added to CodonPhyML to aid in the analysis of results. The following results are returned:

- Site-specific marginal likelihoods, conditioned on the rate class.
- Posterior probability of observing each rate class at all sites.
- Mean posterior rates at all sites.
- Rate class assignments for all sites and a histogram of the assignments.

The likelihood, conditioned on each rate class, for site i , $\Pr(\mathcal{D}_i, \mathcal{T}|\hat{\omega}_j)$, is used in calculating the total likelihood, as shown in the previous chapter. Hence, returning the site-specific likelihoods simply consists of printing these values.

The posterior probability of observing rate class j at site i is given by Bayes formula:

$$\Pr(\hat{\omega}_j|\mathcal{D}_i, \mathcal{T}) = \frac{\Pr(\mathcal{D}_i|\hat{\omega}_j, \mathcal{T}) \Pr(\hat{\omega}_j)}{\sum_{k=1}^n \Pr(\mathcal{D}_i|\hat{\omega}_k, \mathcal{T}) \Pr(\hat{\omega}_k)}$$

These probabilities are approximated by a naïve empirical Bayes (NEB) method as in Nielsen and Yang [1998] and Yang *et al.* [2000]. That is, the maximum likelihood estimates for $\Pr(\omega_j)$ are used in the formula. The mean posterior rates at each site are calculated in the same way, according to the formula:

$$\mathbb{E}(\hat{\omega}|\mathcal{D}_i, \mathcal{T}) = \sum_{j=1}^n \Pr(\hat{\omega}_j|\mathcal{D}_i, \mathcal{T}) \hat{\omega}_j \tag{4.1}$$

where $\hat{\omega}_j$ is used to denote the maximum likelihood estimate for ω_j . In the site-specific likelihoods and posterior probabilities only the different ω rates are used, as they completely characterize the likelihood at a site. From these values the likelihood and posterior probabilities of the different nonsynonymous and synonymous classes may be obtained. On the other hand, the mean posterior rate is calculated for d_N and d_S , where applicable.

Finally, the rate class assignments can be made to the class with the highest likelihood, or the class with the largest posterior probability. It should be noted that when the posterior probability is used to assign sites to rate classes the distribution of rate classes inferred by maximum likelihood no longer holds.

Table 4.1: Statistics of the five datasets used for validation.

Dataset	Sequences	Codons
Japanese encephalitis <i>env</i> gene	23	500
β -globin gene from vertebrates	17	144
<i>Drosophila</i> alcohol dehydrogenase (<i>adh</i>) gene	23	254
Tick-borne flavivirus NS-5 gene	18	342
HIV-1 <i>vif</i> gene	29	192

4.2 Validation

In this section the performance of all the parametric codon models in CodonPhyML is compared to HyPhy on real data. HyPhy was used in the comparisons since codeml does not implement any of the models that were added to CodonPhyML. Although the Constant Rates, Nonsynonymous GDD and Nonsynonymous Γ models have been compared to their counterparts in codeml by Zanetti [2010], the comparisons were only performed for the Goldman-Yang frequency model and only on simulated data. The validation experiments carried out here were done not only to ensure that the models are implemented correctly, but also to gain knowledge on the behaviour of the different model combinations. Lastly, a separate experiment was carried out to investigate how the performance of models with rate heterogeneity scale as more classes are added.

4.2.1 Data

Five datasets of protein coding genes were used in the comparison⁷. A full description of the datasets can be found in [Yang *et al.* 2000].

Yang *et al.* [2000] found that models allowing for rate heterogeneity provide the best fit to all of the datasets. However, they only analyzed models that allow the nonsynonymous rate to vary. They found substantial evidence for diversifying selection in the β -globin and HIV-1 *vif* datasets and evidence for purifying selection on the Japanese encephalitis *env* dataset. In all cases where continuous distributions were used it was found that the distribution was L-shaped, with only a small proportion of the sites under positive selection.

Later analyses found that the dual models are preferred for all the datasets, except the *Drosophila adh* dataset. [Kosakovsky Pond and Frost 2005a]. It was further found that the GBDD model is preferred above the Dual GDD model on the flavivirus NS-5 and Japanese encephalitis *env* datasets [Kosakovsky Pond *et al.* 2010b].

4.2.2 Results

All of the experiments were carried out on Ubuntu 10.04.1 using a Hewlett-Packard Pavillion dv6 Notebook with an Intel Core i5 processor with two 2.40 GHz cores and 4 GB of memory. Each experiment was run 5 times, and in each case only the results from the run with the highest likelihood are reported. Results with the Yap frequency model are only given for CodonPhyML since this model is not available in HyPhy. In all the experiments with rate heterogeneity 3 rate classes were used. This led to a total of 3 rate classes for the Nonsynonymous Γ , Nonsynonymous GDD and GBDD models, but 9 rate classes for the two dual models. Instead of reporting each rate and their respective weights, the coefficient of variation for each rate type is reported as a summary statistic. Unless stated otherwise the tree included with the datasets was used as an initial tree. In all of the experiments only the

⁷Downloaded from <http://www.hyphy.org/pubs/2rates.tgz>

rates and branch lengths were optimized with CodonPhyML and all possible code optimizations were applied. For brevity only the results on the Japanese encephalitis and β -globin datasets are reported here. The results for the remaining 3 datasets can be found in appendix A.

The results for the Constant Rates, Nonsynonymous Γ and Nonsynonymous GDD models are summarized in table 4.2 and the results for the Dual Γ and Dual GDD models are summarized in table 4.3. In all cases the observed nucleotide frequencies were used, and $F3 \times 4$ was used to compute the counts. The HyPhy results were produced with the batch file `dNdSRateAnalysis`. Wherever the Muse-Gaut frequency model was used with HyPhy, the HKY85 nucleotide model was coupled to it.

It can be seen that for the Constant Rates, Nonsynonymous Γ and Dual Γ models the parameter and likelihood values found by CodonPhyML agree very closely to the values found by HyPhy. However, for the GDD models some deviations can be seen. The deviations in the parameter estimates for the Nonsynonymous GDD model are quite small, although the difference between the likelihoods can be significant, especially for the Goldman-Yang model. The deviations are more serious for the Dual GDD model. Although the likelihoods and estimates for κ and $CV(d_N)$ agree, the estimates for $CV(d_S)$ differ in several cases. In most cases where the $CV(d_S)$ estimates differ the $CV(\omega)$ estimates still agree. (See for instance the Japanese encephalitis dataset with Muse-Gaut frequencies or the *Drosophila adh* dataset). Since the likelihood calculation depends on the values of ω , this shows that the optima found by CodonPhyML and HyPhy are very close in these datasets. In three cases there is a sizeable difference between both the $CV(d_S)$ and $CV(\omega)$ estimates. (β -globin dataset with Goldman-Yang frequencies, HIV-1 *vif* dataset with Goldman-Yang frequencies and flavivirus NS-5 dataset with Muse-Gaut frequencies).

The results for the GBDD model are summarized in table 4.4. For this model frequencies were also optimized by maximum-likelihood, using $CF3 \times 4$, which resulted in 9 extra parameters. The HyPhy results were produced with the batch file `dNdSBivariateRateAnalysis`. This batch file only allows the use of the Muse-Gaut frequency model. Hence, no HyPhy results for the Goldman-Yang frequency model are reported. The batch file has two options, ‘Nucleotide Model’ and ‘Codon Model’. When ‘Nucleotide Model’ is selected, initial branch length estimates are calculated using a nucleotide model before proceeding with the optimization of the codon model. This is a poor approximation for some datasets, which leads to differences between the parameter values estimated from ‘Nucleotide Model’ and ‘Codon Model’ (personal communication, Sergei Kosakovsky Pond). Because optimization in CodonPhyML is similar to ‘Codon Model’, and because it is believed to give a better approximation, it was used throughout⁸. In all experiments the HKY85 nucleotide model was coupled to the Muse-Gaut frequency model.

Although there are differences between the parameter estimates, the agreement between CodonPhyML and HyPhy is in general higher than with the Dual GDD model. In particular the likelihood values are nearly identical for both programs on all of the datasets. Once again, only the estimates for $CV(d_S)$ and $CV(\omega)$ show any serious discrepancies. The most pronounced differences are for $CV(d_S)$ in the Japanese Encephalitis and *Drosophila adh* datasets.

The results for the Constant Rates + ψ model are summarized in table 4.5. For these experiments the preferred codons found by Zhou *et al.* [2010] for Yeast were used. It should be noted that since this is only a test for the correctness of the implementation and not an investigation into selection on the codon usage, the actual codons assigned to the preferred and unpreferred classes are arbitrary. The HyPhy results were produced by a script provided by the authors of [Zhou *et al.* 2010]⁹ The script uses the GTR nucleotide model coupled with the Goldman-Yang frequency model. For these experiments it was modified to use the HKY85 nucleotide model. In all five datasets the estimates by CodonPhyML and HyPhy were in close agreement, showing that the two implementations are estimating the same model. Since adding the ψ parameter to a model with rate heterogeneity is not conceptually any different to adding it to the Constant Rates model, the performance of these models with the ψ parameter was not validated.

It should be noted that the two models that show the biggest disagreement between CodonPhyML and HyPhy, the Dual GDD and GBDD models, are also the two models with the most free parameters (figure 4.1 and 4.2). It is by no means surprising that these models would be more unstable, since

⁸Initially, optimization under ‘Codon Model’ lead to vastly different results between CodonPhyML and HyPhy. This discrepancy was brought to the attention of the authors of HyPhy, and was found to be caused by a bug in HyPhy, which has subsequently been fixed.

⁹Downloaded from <http://openwetware.org/images/5/52/Zhou-Gu-Wilke-synonymous-selection.zip>

Table 4.2: Validation results for the Constant Rates, Nonsynonymous Γ and Nonsynonymous GDD models.

	Constant Rates			Nonsynonymous Γ			Nonsynonymous GDD		
	$\ell(\hat{\theta}, \mathcal{D}, \mathcal{T})$	κ	ω	$\ell(\hat{\theta}, \mathcal{D}, \mathcal{T})$	κ	$CV(\omega)$	$\ell(\hat{\theta}, \mathcal{D}, \mathcal{T})$	κ	$CV(\omega)$
Japanese encephalitis									
<hr/>									
Goldman-Yang									
CodonPhyML	-6886.1641	9.5238	0.0507	-6845.3819	9.5555	2.0170	-6806.8535	9.1027	1.4115
HyPhy	-6886.1641	9.5252	0.0507	-6845.3010	9.5689	2.0968	-6853.9516	9.5658	1.4142
Muse-Gaut									
CodonPhyML	-6839.6735	9.0816	0.0489	-6797.1614	9.0868	2.0904	-6807.4700	8.8846	1.4115
HyPhy	-6839.6734	9.0828	0.0489	-6797.1778	9.0888	2.0956	-6806.8419	9.1032	1.4142
Yap									
CodonPhyML	-6840.5962	8.8646	0.0482	-6797.3866	8.8741	2.1168	-6807.4700	8.8846	1.4115
<hr/>									
β -globin									
<hr/>									
Goldman-Yang									
CodonPhyML	-3815.5142	2.0704	0.2369	-3687.0641	2.1451	1.4116	-3684.3850	1.9914	0.9613
HyPhy	-3815.5141	2.0710	0.2369	-3687.0639	2.1449	1.4109	-3702.2154	2.0759	0.9948
Muse-Gaut									
CodonPhyML	-3784.2302	2.0085	0.2673	-3671.8135	2.0346	1.2195	-3685.4477	1.9701	0.9640
HyPhy	-3784.2302	2.0092	0.2673	-3671.8134	2.0347	1.2193	-3684.3848	1.9910	0.9612
Yap									
CodonPhyML	-3786.1733	1.9895	0.2644	-3673.0290	2.0173	1.2233	-3685.4477	1.9701	0.9640

Table 4.3: Validation results for the Dual Γ and Dual GDD models.

	Dual Γ				Dual GDD					
	$\ell(\hat{\theta}, \mathcal{D}, \mathcal{T})$	κ	$CV(d_N)$	$CV(d_S)$	$CV(\omega)$	$\ell(\hat{\theta}, \mathcal{D}, \mathcal{T})$	κ	$CV(d_N)$	$CV(d_S)$	$CV(\omega)$
Japanese encephalitis										
<hr/>										
Goldman-Yang										
CodonPhyML	-6812.5979	9.3668	1.4113	0.6038	1.7889	-6800.1875	9.3469	2.0239	0.7325	2.6933
HyPhy	-6812.5904	9.3661	1.4142	0.6037	1.7920	-6800.0833	9.3621	2.1167	0.7661	2.7530
Muse-Gaut										
CodonPhyML	-6785.1753	9.0867	1.4114	0.5135	1.6865	-6773.4328	9.0670	2.0955	0.4152	9.2039
HyPhy	-6785.1626	9.0861	1.4142	0.5135	1.6894	-6773.4327	9.0655	2.0955	0.5774	9.2057
Yap										
CodonPhyML	-6785.2171	8.9383	1.4114	0.5157	1.6887	-6773.0813	8.9428	2.1226	0.4319	6.4333
<hr/>										
β -globin										
<hr/>										
Goldman-Yang										
CodonPhyML	-3687.4328	1.9805	1.0013	0.6252	1.3693	-3665.9011	1.9569	1.4133	0.4182	15.5127
HyPhy	-3687.4324	1.9799	1.0015	0.6254	1.3697	-3666.6498	1.9632	1.4118	1.0478	1.5752
Muse-Gaut										
CodonPhyML	-3675.0207	1.9419	0.9664	0.5484	1.2522	-3659.0524	1.9272	1.2223	2.0315	1.3183
HyPhy	-3675.0205	1.9416	0.9664	0.5484	1.2522	-3659.2987	1.9394	1.2282	0.9893	1.4137
Yap										
CodonPhyML	-3676.0550	1.9240	0.9691	0.5492	1.2558	-3660.2452	1.9143	1.2250	1.9330	1.3108

Table 4.4: Validation results for the GBDD model.

	$\ell(\hat{\theta}, \mathcal{D}, \mathcal{T})$	κ	$CV(d_N)$	$CV(d_S)$	$CV(\omega)$		
<hr/>							
Goldman-Yang						Japanese encephalitis	
CodonPhyML	-6765.7085	8.8620	2.0753	0.4986	2.1262		
Muse-Gaut							
CodonPhyML	-6761.9154	9.0731	2.0948	0.2733	2.8568		
HyPhy	-6761.9155	9.0711	2.1127	0.5307	2.5919		
Yap							
CodonPhyML	-6763.1820	8.9283	2.1297	0.5007	2.1811		
<hr/>							
Goldman-Yang							β -globin
CodonPhyML	-3613.9822	1.7176	1.2539	0.0723	1.2910		
Muse-Gaut							
CodonPhyML	-3632.9037	1.8368	1.2951	0.1048	1.3512		
HyPhy	-3632.9038	1.8367	1.2913	0.1055	1.3165		
Yap							
CodonPhyML	-3633.8111	1.8192	1.2969	0.1104	1.3560		
<hr/>							

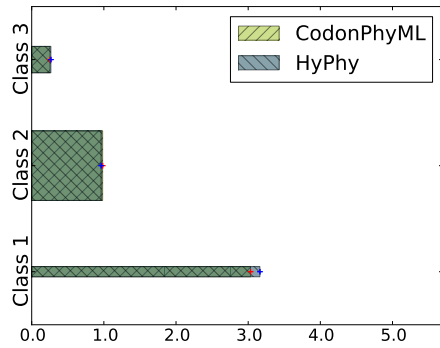
their likelihood surfaces are expected to be more complex. Hence, even minor differences in the implementations could lead to different optima being found. Furthermore, CodonPhyML and HyPhy use different numerical optimization schemes, which may also lead to different optima being found. The differences could also be due to the peculiarities of the datasets. There was a disagreement on the flavivirus NS-5 dataset for the Dual GDD model. This dataset is the most conserved of all the datasets, hence it may not carry enough information for reliable estimates. Furthermore, diversifying selection has been detected on the β -globin and HIV-1 *vif* datasets, which could explain the higher variation in the estimates.

Since most of the variation between CodonPhyML and HyPhy occurred in the estimates for $CV(d_S)$, the synonymous rates that were estimated for these two models are compared in figure 4.3. In the figure the lengths of the bars represent the estimated rates, while the width of the bars are proportional to the weights assigned to the respective rate classes. The plotted values are the median estimates from 5 runs. A trimmed sample (with the smallest and largest values removed) was used to plot the variation in the rates. The same trimming was applied in plotting the variation on the weights. This was done due to the presence of outliers. In most cases, runs with an outlier found one rate that was overestimated, and then assigned a weight of almost zero to it. In all of the plots the rates are ordered from the largest rate to the smallest rate.

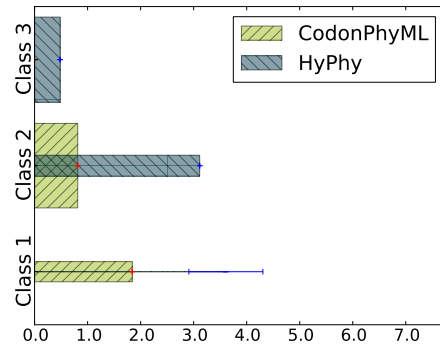
On the Dual GDD model the only serious difference between the estimated rates are for the β -globin dataset with the Goldman-Yang frequency model. It can be seen that both HyPhy and CodonPhyML assign a weight of almost zero to one of the rates. Not surprisingly, this rate has a high variation. On the GBDD model the estimates differ more between the implementations, with only the β -globin and HIV-1 *vif* datasets finding similar estimates. On the datasets where the estimates differ it can be seen that the CodonPhyML estimates show a much larger variation than the HyPhy estimates. The variation is not surprising, since this model has a lot of freedom and is known to be quite unstable.

In all of the results presented above it can be seen that there is quite a large difference between the

Dual GDD model with Goldman-Yang frequencies

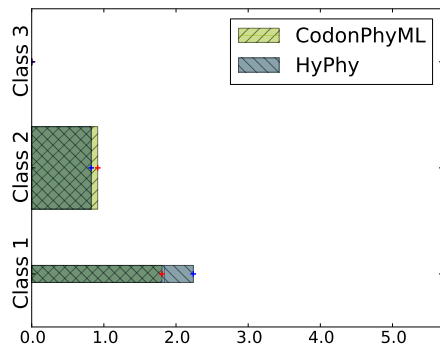


(a) Japanese encephalitis

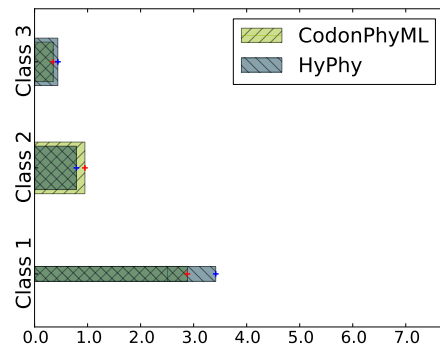


(b) β -globin

Dual GDD model with Muse-Gaut frequencies

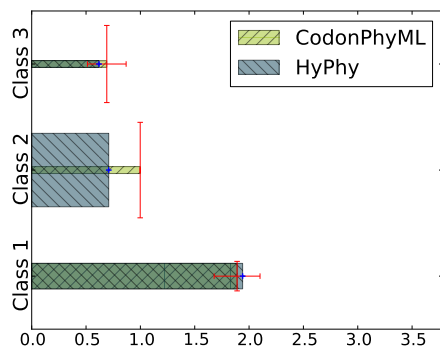


(c) Japanese encephalitis

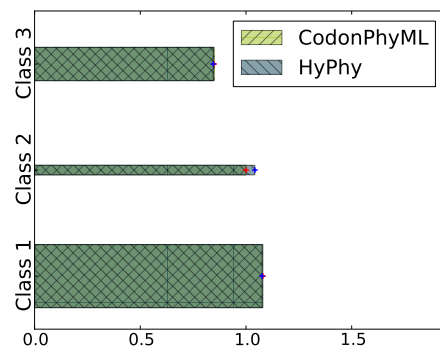


(d) β -globin

GBDD model with Muse-Gaut frequencies



(e) Japanese encephalitis



(f) β -globin

Figure 4.3: Comparison of the synonymous rates found using CodonPhyML and HyPhy with the Dual GDD and GBDD models coupled with different frequency models. The width of the bars are proportional to the weight assigned to each rate. The values plotted are for the median rate from 5 runs. The error bars for CodonPhyML and HyPhy are respectively red and blue and were produced from a trimmed sample. (In each case the smallest and largest values were removed).

Table 4.5: Validation results for the Constant Rates + ψ model.

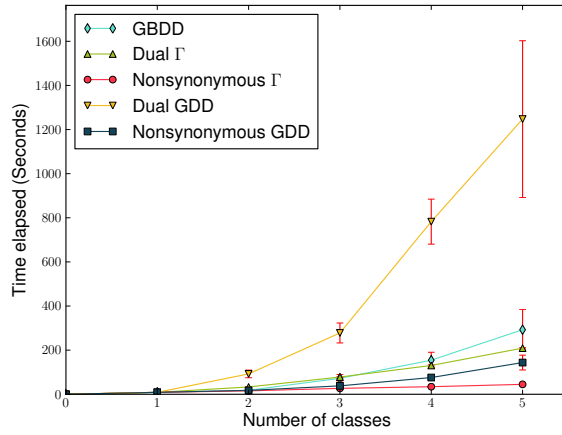
	$\ell(\hat{\theta}, \mathcal{D}, \mathcal{T})$	κ	ω	ψ	
<hr/>					Japanese encephalitis
Goldman-Yang					
CodonPhyML	-6879.6673	10.2810	1.2944	0.0578	
HyPhy	-6879.6678	10.2513	1.2928	0.0578	
Muse-Gaut					
CodonPhyML	-6836.4898	9.4624	1.1948	0.0536	
Yap					
CodonPhyML	-6839.1188	9.1077	1.1290	0.0514	
<hr/>					
Goldman-Yang					β -globin
CodonPhyML	-3815.4085	2.0709	1.0624	0.2454	
HyPhy	-3815.4086	2.0726	1.0650	0.2459	
Muse-Gaut					
CodonPhyML	-3782.9240	2.0039	1.2325	0.3012	
Yap					
CodonPhyML	-3785.0206	1.9850	1.2172	0.2958	
<hr/>					

different frequency models. While the parameter estimates for the Yap model are quite close to the estimates for the Muse-Gaut model, parameter estimates for the Goldman-Yang model tends to deviate more from the other models. It may perhaps seem surprising that the Goldman-Yang model often has a considerably worse likelihood, as it is predicted to find a closer fit to the data [Yap *et al.* 2010]. This behaviour is due to the fact that the frequencies were not optimized by maximum-likelihood, but counted from the data. When the frequency parameters were optimized by maximum-likelihood the Goldman-Yang model achieved a better likelihood than either of the other two models (results not shown). Although this does not hold in all cases it can be observed in table 4.4 for the β -globin dataset. (Recall that frequency parameters were optimized for the GBDD model).

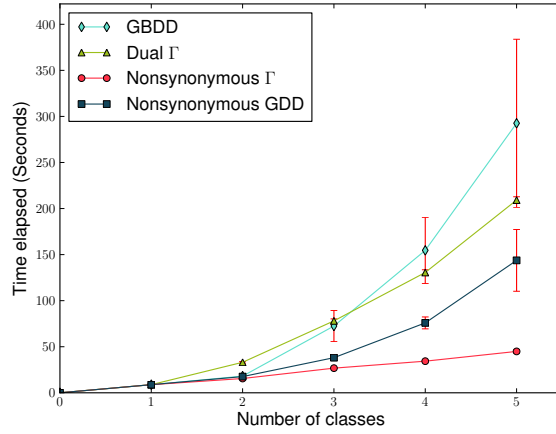
A final experiment was carried out to investigate how the performance of CodonPhyML scales on models with rate heterogeneity as more classes are added. This experiment was only performed on the β -globin dataset. The results are shown in figure 4.4. Each plotted value is the mean value of 5 runs. As expected, the Dual GDD model requires much more time to find an optimum than any of the other models. Unlike the GDD models, the number of free parameters for the Γ models does not increase as more classes are added. However, the time taken still increases, as the likelihood calculation requires the construction of more matrices. It is interesting to observe that for less than 4 classes the GBDD model is faster than the Dual Γ model. It is also evident here that the GDD models have more variation with regards to the time needed to find an optimum. As with the parameter values, this is due to the models having more free parameters. Lastly, the memory usage of the dual models increases quadratically, while the usage by the other models increases linearly, as expected.

4.3 Discussion

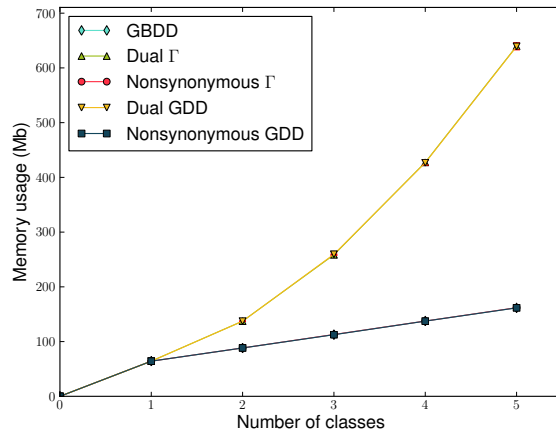
In this chapter a number of additions to CodonPhyML are presented. CodonPhyML now has a wide range of parametric codon models with rate heterogeneity available. Furthermore, selection on codon



(a)



(b)



(c)

Figure 4.4: The performance of the different models of rate heterogeneity as the number of classes are increased. In (a) the time taken by the different models is shown. In (b) the time for the Dual GDD model is not plotted. In (c) the memory usages of the different models are shown. All values plotted are the mean values from 5 runs.

usage can also be investigated. Importantly, CodonPhyML is still able to consistently outperform other software packages. In all of the experiments presented above CodonPhyML found an optimum faster than HyPhy, and in some cases did so in less than a quarter of the time. However, this tradeoff comes at a price, CodonPhyML is more memory-intensive than HyPhy, often using more than twice as much memory. It should also be kept in mind that CodonPhyML is less reliable than HyPhy when using the GBDD model and often shows a larger variation between runs.

Instead of regarding the disagreement between CodonPhyML and HyPhy as one package being wrong and the other right, it should be seen as two slightly different optima being found by the two packages. The likelihood surfaces of these models are extremely complicated and it is difficult to tell which solution is actually the best fitting solution when the likelihood values are close to each other.

A number of further additions could be made to CodonPhyML to improve its utility. In particular, CodonPhyML does not currently implement any models with branch specific rate parameters. Furthermore, the naïve empirical Bayes estimation in the post-processing step should be replaced with a Bayes empirical Bayes (BEB) estimation as described in [Yang *et al.* 2005] in order to account for uncertainties in the maximum-likelihood estimations of parameter values. This is especially relevant since the validation results show that some models have a large variation in the estimated parameters. A further post-processing step should also calculate Bayes factors, which account better for the prior distributions. In particular, when looking for evidence of selection, Bayes factors give the relative and not the absolute effect size [Kosakovsky Pond and Frost 2005a]. Finally, it would also be beneficial to return the probabilistic ancestral sequences inferred at each location in the tree.

Chapter 5

Data description and model selection

Before tests for correlations between evolutionary rate variations, synonymous codon usages and protein secondary structures can be carried out the data have to be collected from somewhere. Of these quantities the codon usage statistics are the easiest to collect. The only thing needed is the sequence itself and there is no shortage of databases with DNA sequences. Secondary structure assignments are scarcer, but can be readily extracted from the PDB. It is more difficult to obtain site-specific assignments of the evolutionary rates. In order to estimate the evolutionary rates a multiple sequence alignment of orthologs is required. Furthermore, the rates that are estimated are dependent on the model and the other sequences in the alignment. Since the rates are conditioned on the set of sequences and since there is no consensus on which model is the correct model for any set of sequences, these rates cannot simply be extracted from a database, but must be obtained through simulations.

In this chapter the strategy that was followed to obtain the data that is analyzed in the next chapter is discussed. The chapter starts with some background on model selection strategies, before giving the procedure that was used to extract sets of orthologs with protein secondary structure assignments. This is followed by a description of the model selection strategy that was used to select the best-fitting model for every group of orthologs, interspersed with summaries of the intermediate results. The aim was to find those groups with evidence of synonymous variation or evidence of selection on the synonymous codon usage.

Aside from producing the data for the next chapter the datasets discussed in this chapter could be useful in subsequent investigations. Fitting the models was by far the most time-consuming part of the project, requiring about a month's computation on between 10 and 30 computers at a time. This process produced a huge amount of data that are not touched upon in this report and could be investigated in future projects.

5.1 Background

It is unlikely that any of the models described in chapter 3 is the true model underlying molecular evolution. However, it is obvious that some models will give a better description for a dataset than others. It is also clear that the best model for one dataset is not necessarily the best model for another dataset. In order to decide which model gives the best fit, a model selection strategy needs to be followed. The most straightforward method would be to use maximum-likelihood. However, the model with the best likelihood is not necessarily the best model. The likelihood increase may not be significant when the extra complexity of the model is considered, due to the possibility of overfitting. This highlights the need for a more rigorous approach to assess the fit of a model. Furthermore, fitting all the models to a dataset may be time consuming and wasteful, especially if some of the models are never selected. Burnham and Anderson [2002] recommends using logical arguments to reduce the number of candidate models prior to model selection. In this section a brief overview of model selection strategies for nested and non-nested models are given. Further, some points about the sample size with relation to stochastic codon models are discussed. Finally, if it is not clear which model provides

the best fit, it is possible to perform a multimodel inference, using model averaged predictions from all of the candidate models. Model averaging is not discussed here.

5.1.1 Nested models

Suppose two models, a null model (H_0) and an alternative model (H_1) have been used to describe a dataset, \mathcal{D} . The models, H_0 and H_1 respectively have parameter vectors $\hat{\theta}_1$ and $\hat{\theta}_0$ and likelihoods $\mathcal{L}(\hat{\theta}_1|\mathcal{D})$ and $\mathcal{L}(\hat{\theta}_0|\mathcal{D})$.

If the models are nested, an LRT can be used to determine if the increase in likelihood for H_1 is significant. The two models are nested if fixing some parameters in H_1 makes it equivalent to H_0 . Then $\mathcal{L}(\hat{\theta}_1|\mathcal{D}) \geq \mathcal{L}(\hat{\theta}_0|\mathcal{D})$ should hold, since H_0 is incorporated within H_1 . The test statistic for an LRT is:

$$2 \left[\frac{\mathcal{L}(\hat{\theta}_1|\mathcal{D})}{\mathcal{L}(\hat{\theta}_0|\mathcal{D})} \right]$$

or

$$2\ell(\hat{\theta}_1|\mathcal{D}) - 2\ell(\hat{\theta}_0|\mathcal{D}) = 2\Delta\ell \tag{5.1}$$

if the log-likelihoods are used. If H_0 holds, then under some regularity conditions, the test statistic for a large sample will follow a χ^2 distribution with degrees of freedom equal to the difference in the number of free parameters between the models [Anisimova *et al.* 2001]. The regularity conditions are violated when parameters in H_1 have to be fixed to the boundary values of the parameter space in order to make the model equivalent to H_0 . If all conditions are satisfied, then H_0 can be rejected in favour of H_1 if $2\Delta\ell > \chi_n^2(\alpha)$ where n is the difference in degrees of freedom between the models and α the significance level.

When the best fitting model is to be selected from a set of candidate models several different strategies may be followed. The simplest strategy is to fit all the models on the data and then perform all possible LRTs. Such a comparison leads to a large number of tests, which can be difficult to analyze. A hierarchical approach involves less LRTs and it is often not necessary to fit all the models on the data. However, it should be noted that a hierarchical approach to model selection may find only a local optimum [Bao *et al.* 2007]. In forward selection, the simplest model is fit first. LRTs are then conducted for increasingly more complicated models in a step-by-step fashion until the increase in complexity does not lead to a significant improvement in the likelihood. Backward elimination works in the opposite direction, starting with the most complicated model and then progressing to simpler models. While forward selection tests if adding parameters increases the fit, backward elimination removes parameters that do not result in a significant likelihood increase. In most cases forward selection will be the least costly method, since it starts with the simplest model.

5.1.2 Non-nested models

If the models are not nested the empirical distribution of likelihoods can be obtained from Monte Carlo simulations under H_0 . However, this is extremely time consuming and will not be examined here. Instead, a model selection criterion can be defined, consisting of the log-likelihood and a penalty term to correct for the model bias. Two of the most widely used criteria for model selection are discussed below. The derivation of these criteria are complicated and not discussed here. Furthermore, several other related criteria exist. The interested reader is referred to [Burnham and Anderson 2002]. Regardless of which criterion is used, every model has to be fit to the data in order to make a decision. Finally, note that these criteria may also be used for nested models.

The Akaike Information Criterion (AIC) is based on an information-theoretic approach. The assumption is made that there is no true model and instead it tries to select the model that gives the closest approximation to reality [Burnham and Anderson 2002]. The AIC is an approximation to the Kullback-Leibler information of a model, which measures the amount of information lost when using the model to approximate reality [Burnham and Anderson 2002]. The bias correction in the AIC is dependent on the number of free parameters in the model, k :

$$\text{AIC} = -2\ell + 2k \tag{5.2}$$

When using the AIC for model selection, the best fitting model is simply the model that has the smallest AIC. (Note that the model with the smallest AIC is not necessarily a good approximation to reality, but merely a better approximation than the other candidate models).

It is possible for a model to not have the lowest AIC, but still be reasonably well supported. The important criterion for measuring support is not the absolute size of the AIC, but the relative differences between the set of candidate models. In this we follow Burnham and Anderson [2002] who assign varying levels of support to model i based on the value of Δ_i , the difference between AIC_i and AIC_{\min} , where AIC_{\min} is the AIC value of the best-fitting model. The levels of support we assign are given in table 5.1. In accordance to this table only models with $\Delta_i \leq 2$ are considered to have significant support.

Models are more sensitive to overfitting when the sample size is not much larger than the number of free parameters. In these cases the AIC is not conservative enough and a stronger bias correction is needed. This is done by using a second order approximation to the Kullback-Leibler information, called the corrected AIC (AIC_c):

$$\text{AIC}_c = \text{AIC} + \frac{2k(k+1)}{n-k-1} = -2\ell + 2k + \frac{2k(k+1)}{n-k-1} \tag{5.3}$$

where n is the sample size. In the AIC the likelihood is simply penalized by the number of free parameters in the model. In the AIC_c this penalty also depends on the sample size. AIC_c should be used in cases where $\frac{n}{k} < 40$. From the definition it is clear that the AIC_c will tend to pick the same or a simpler model than the AIC.

The Bayesian Information Criterion (BIC) approaches model selection from a Bayesian perspective. Its derivation is motivated by the assumption that a true model exists and that it is one of the candidate models [Burnham and Anderson 2002]. The BIC is defined as:

$$\text{BIC} = -2\ell + 2k \log(n) \tag{5.4}$$

Implicit in the definition of the BIC is the assumption that the true model remains fixed as the sample size is increased. Hence, as the sample size approaches infinity the posterior probability of a model becomes 1, if the model is the true model [Burnham and Anderson 2002]. Model selection with the BIC is performed in exactly the same way as with the AIC. Note that the BIC does *not* approximate the relative Kullback-Leibler distance [Burnham and Anderson 2002].

5.1.3 Sample Size

It is not immediately clear which value should be used as the sample size for stochastic codon models. The definition of the sample size is poorly understood and is also dependent on the quantity of interest [Posada and Buckley 2004]. Here n is set equal to the number of sites in the alignment. This is equivalent to the traditional definition of the sample size for maximum-likelihood methods as the

Table 5.1: Assigning support to models using the AIC.

AIC Difference	Interpretation
$\Delta_i \leq 2$	Strong support
$2 < \Delta_i \leq 4$	Moderate support
$4 < \Delta_i \leq 7$	Less support
$7 < \Delta_i \leq 10$	Almost no support
$10 < \Delta_i$	No support

number of independent observations. (That is, the number of terms in the sum that is used to compute the log-likelihood). However, it is clear that the likelihood depends also on the number of sequences in the alignment and not just on the length of the sequences. Delpont *et al.* [2010] and Kosakovsky Pond *et al.* [2010b] set the sample size to the total number of codons in the alignment. We refer to the version of the AIC_c using this definition of the sample size as AIC'_c . In reality, the sample size lies somewhere in between the number of sites and the number of characters [Posada and Buckley 2004]. We use the number of sites, which is an underestimate, as this leads to a more rigorous model selection procedure, as shown below.

The AIC_c can be written as:

$$AIC_c = -2\ell + 2 \left(\frac{nk}{n-k-1} \right) \quad (5.5)$$

where it can be seen that the penalty term depends on the product of the sample size and the number of free parameters in the numerator and on the difference in the denominator. In what follows it is shown that using AIC'_c leads to a less preferable condition when branch lengths are among the free parameters. Intuitively this makes sense because the number of sequences have already been accounted for in the number of free parameters. In such cases, the number of free parameters, k can be rewritten as $2t+r-3$ where t is the number of sequences in the alignment and r is the remaining number of free parameters (usually r consists of the parameters of the rate matrix). Letting the sample size equal the number of sites in the alignment (sequence length) and setting the number of sites to n we obtain:

$$AIC_c = -2\ell + \frac{2n(2t+r-3)}{n-2t-r+2} \quad (5.6)$$

In contrast, setting the sample size equal to the number of codons (due to gaps in the alignment the number of codons is only roughly equal to nt):

$$\begin{aligned} AIC'_c &\simeq -2\ell + \frac{2nt(2t+r-3)}{nt-2t-r+2} \\ &= -2\ell + \frac{2n(2t+r-3)}{n-2-\frac{r}{t}+\frac{2}{t}} \end{aligned} \quad (5.7)$$

We note that in general (and in all the models considered here), $t > r$ and $s \gg 1$. We conclude by noting that for AIC_c the penalty term is proportional to $\frac{sk}{s-k}$, but for AIC'_c it is proportional to $\frac{sk}{s}$. Where the denominator is significantly reduced in size by the number of parameters with AIC_c , for AIC'_c , this is not the case, and the denominator is not as affected by the number of parameters.

The use of the AIC_c depends on the ratio of the sample size and the number of free parameters, and not the sample size alone. As the data size increases the correction in AIC_c tends toward 0. With AIC'_c the correction term is much smaller and disappears much faster as the number of sites are increased. In some cases this could lead to the bias correction not being strong enough. (In fact, if the number of sequences and the number of parameters are held constant, the second order term only makes a noticeable difference over a very small interval).

On the other hand, as noted by Delpont *et al.* [2010], defining the sample size as the number of codons does indeed lead to the BIC being more conservative. However, we still recommend using the number of sites as the sample size, for consistency and because the number of sequences are already taken into account in the number of free parameters.

5.2 Data Extraction and Preparation

This section details the procedure that was followed to collect the data used in the next chapter to look for relationships between rates, frequencies and structures. Several researchers have reported differences in the relationship between the codon usage and secondary structure in eukaryotes and prokaryotes [Gu *et al.* 2004, Oresic and Shalloway 1998, Tao and Dafu 1998]. Therefore, it was decided to investigate effects for eukaryotes and prokaryotes separately. The datasets themselves consist of groups of orthologs, where for each group secondary structure assignments are available for the model

organism. Furthermore, for consistency the requirement was made that all the groups in a dataset should consist of sequences from the same set of organisms. For the prokaryote dataset *E. coli* was chosen as the model organism and *Homo sapiens* was chosen for the Eukaryote dataset.

The data collection and preparation pipeline discussed below was semi-automated using Python scripts, including Biopython libraries [Cock *et al.* 2009].

5.2.1 Extraction of orthologous groups

The same protocol was followed for the collection of both datasets. The search was started by performing a query for protein structures of the model organism in the Protein Structure Database¹ (PDB). A 90% identity cutoff on sequence similarity was imposed on all the structures returned from this search. This was done to remove paralogs and multiple structures for the same sequence. Sets of orthologs were extracted from the Orthologous MAtrix² (OMA) database. First, Uniprot identifiers for the structure sequences were extracted and mapped to OMA groups. An OMA group is a set of sequences that are orthologous to each other [Altenhoff *et al.* 2011, Schneider *et al.* 2007]. Structures that did not map to any OMA groups were discarded. For each dataset, a set of closely related organisms was then selected such that the number of OMA groups containing all of the selected organisms was maximized. In the prokaryote dataset 20 strains of *E. coli* were selected, and in the eukaryote dataset 15 placental mammals were selected. Henceforth the datasets are referred to as the *E. coli* and Mammalian datasets. Details of the organisms selected and their phylogenetic relationships are given in figure 5.1. After this process the *E. coli* dataset was reduced from an initial 9,613 structures to 826 OMA groups. The Mammalian dataset was reduced from 20,896 structures to 985 OMA groups.

Protein and cDNA sequences for all of the selected organisms in the remaining OMA groups in both datasets were then extracted from the OMA database. Groups where the shortest sequence was less than 100 codons in length and groups with less than 5 unique cDNA sequences were excluded. For some of the groups in the *E. coli* dataset the coding sequences stored in the OMA database are incomplete. These groups were subsequently excluded as well. Secondary structure assignments were then extracted from the PDB for all the sequences remaining in the datasets. Three classes of secondary structure are assigned, helices, sheets and loops. At times, secondary structure assignments are not available for all sites in a sequence. Groups where no structure assignments could be found for the model organism were excluded. (In theory, at least, one structure should be available for the model organism in all of the groups, since the search was started with the results of a query on the PDB. The causes behind this phenomenon were not investigated, as this only happened for a few groups. It is probable that a PDB query on an organism returns some structures from closely related organisms as well). After these steps 542 groups remained in the *E. coli* dataset and 912 groups in the Mammalian dataset. The datasets are summarized in table 5.2.

Sequences in the Mammalian dataset are more diverse, with each group having at least 13 unique sequences, compared to at least 6 unique sequences for every group in the *E. coli* dataset. The sequences in the Mammalian dataset are also on average longer, with the average length of sequences for *Homo sapiens* measuring 614.59 codons and the average length for *E. coli K12* measuring 330.98 codons. Secondary structure assignments are available for 161,548 of the total 179,392 codons (or 90.05%) for *E. coli K12*. In contrast to this, secondary structure assignments for *Homo sapiens* are only available for 234,951 of the 560,506 codons (41.92%). None of the other organisms in either of the datasets have a significant amount of structure assignments available.

The data collection procedure described above was also repeated using *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana* as model organisms. However, these datasets were much smaller, due to the model organisms not being as well represented in the PDB as *E. coli K12* and *Homo sapiens*. Due to the small size of these datasets they were not analyzed further. All PDB searches and sequence downloads were performed on 1 June 2011 and the May 2011 release of the OMA database was used.

¹Available at <http://www.pdb.org>

²Available at <http://www.omabrowser.net>

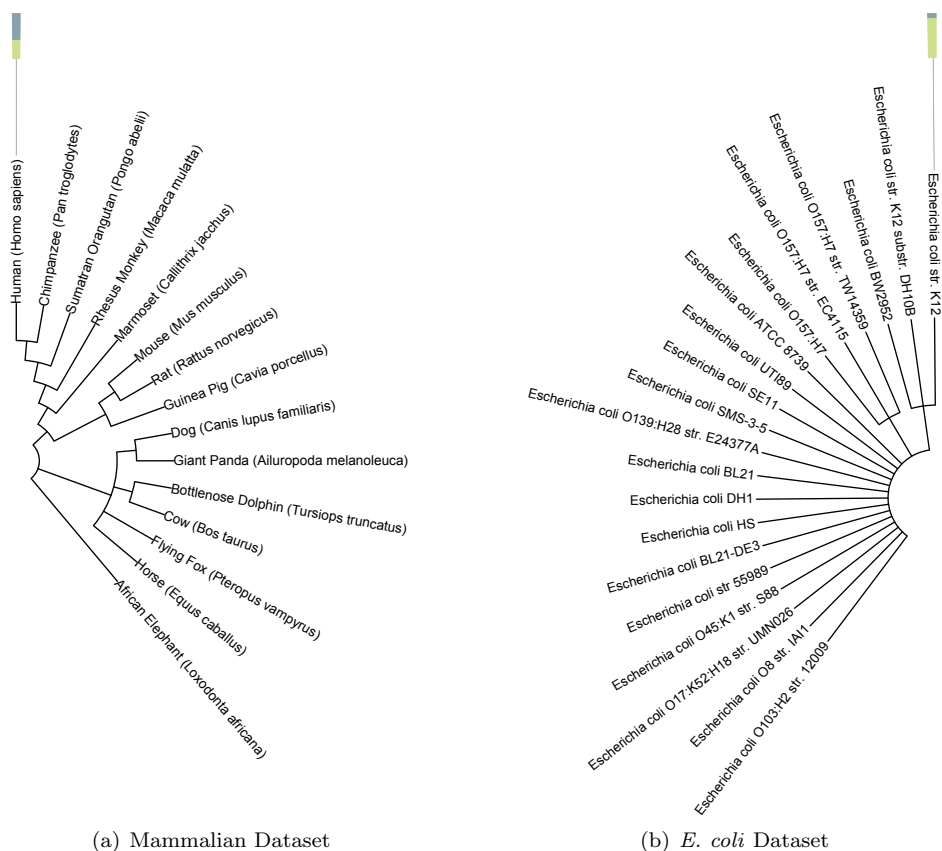


Figure 5.1: NCBI phylogenies of the two datasets. The bar indicates the number of amino acids with secondary structure assignments of the total number of amino acids, for the model organisms. The phylogenies were produced using the Interactive Tree Of Life (iTOL) [Letunic and Bork 2011; 2007].

5.2.2 Pre-processing of orthologous groups

The amino acid sequences of every group in the two datasets were aligned using Mafft L-INS-i [Katoh and Toh 2008] (version 6.847b). Columns in the alignments where more than 75% of the sequences contain a gap were excluded from the alignment. The corresponding nucleotide alignments were then produced by back-translating the amino acid alignments. This strategy is advocated by Dessimoz and Gil [2010] who found that it nearly always produces better results than directly aligning nucleotide sequences. Mafft L-INS-i was chosen as it is consistently ranked as one of the most accurate alignment programs, while also being one of the fastest [Thompson *et al.* 2011]. Although there are some programs that produce marginally better alignments on benchmark sets, the time taken to compute these alignments is much longer. Moreover, Dessimoz and Gil [2010] mention that the differences in amino acid alignments produced by different programs are usually minor, which motivated the decision to use the fastest program.

In theory, the gene tree for each OMA group should have the same topology as the species tree [Altenhoff *et al.* 2011]. This recommends using the trees extracted from the NCBI taxonomy as the topology for each group. However, at least for the *E. coli* dataset this assumption is not expected to hold, due to the closeness of the different strains and an abundance of lateral gene transfers. In fact, it is believed that different *E. coli* strains form a continuum. It has been shown that using different features for classification results in completely different phylogenies [Lukjancenko *et al.* 2010]. This is thought to be caused by the large differences in the genes present in different *E. coli* strains.

Instead of using species trees, a candidate phylogeny was found for each group by fitting the Constant Rates model under the Goldman-Yang frequency model to the nucleotide alignment for each group. The model was fitted using CodonPhyML, optimizing the topology, branch lengths and rates. BioNJ was used for the computation of the initial distance trees. As expected, the topologies show a

Table 5.2: Statistics of the two datasets.

Dataset	OMA Groups	Sequences/Group	Model Organism	Codons with Structure Assignments	Total nr. of codons
<i>E. coli</i>	542	20	<i>E. coli</i> K12	161,549	179,392
Mammalian	912	15	<i>Homo sapiens</i>	234,951	560,506

lot more variation for the *E. coli* dataset than for the Mammalian dataset, where the topologies are in general close to the species tree. In defense of this approach Yang *et al.* [2000] stated that the tree topology does not have a large effect on the estimates of rate parameters and that any reasonably good tree will give similar results to the best (or true) topology.

5.2.3 Extraction of preferred codons

To determine the codon bias in an organism computationally, a large database of coding sequences for the organism is needed. This information is available in the form of the Codon Usage Database³ (CUTG) [Nakamura *et al.* 2000]. However, the CUTG has not been updated since September 2007. Consequently, the CUTG contains only a few or no sequences for many of the organisms in the two datasets discussed above. Instead of using the CUTG, all the available coding sequences for an organism in the OMA database were used.

A similar procedure to the one described in [Zhou *et al.* 2010] was used to extract the preferred codons for an organism. For each organism all of its coding sequences in the OMA database were extracted. The \hat{N}'_c statistic was then computed for each sequence using ENCPPrime⁴ [Novembre 2002]. Due to a bug in the program, only sequences containing all four nucleotides were used. A set of highly biased sequences and a set of unbiased sequences for each organism was created by extracting the 5% of sequences with the respectively lowest and highest \hat{N}'_c values. For each codon in each sequence in the highly biased set the X and X^2 values were calculated as:

$$X = \frac{(f_{obs} - f_{exp})}{f_{exp}} \quad (5.8)$$

$$X^2 = \frac{(f_{obs} - f_{exp})^2}{f_{exp}} \quad (5.9)$$

where f_{obs} is the observed count for the codon in the sequence and f_{exp} the expected count. The expected counts for a codon in a sequence were calculated from the expected synonymous frequency of a codon. Instead of setting all the expected synonymous frequencies for a given codon family to equal values, nucleotide biases were taken into account by using the observed synonymous frequency of a codon in the set of unbiased sequences.

The χ^2 statistic for each codon was obtained by summing all of the X^2 values for that codon in all of the sequences in the set of highly biased sequences. A χ^2 -test was then performed at a significance level of 0.05 to look for differential codon usage. Where a significant effect was found, the X values were added together. If the value was positive the codon was marked as preferred. For each dataset there is a high correspondence in the preferred codons between organisms. If a codon was preferred

³Available at <http://www.kazusa.or.jp/codon/>

⁴Available at <http://www.eeb.ucla.edu/Faculty/Novembre/software/software.html>

Table 5.3: Preferred codons selected for the two datasets.

	Mammalian Dataset	<i>E. coli</i> Dataset
Alanine	GCC	GCG
Arginine	CGC	CGC
	CGG	CGT
Asparagine	AAC	AAC
Aspartic Acid	GAC	GAC
Cysteine	TGC	
Glutamic Acid	GAG	
Glutamine	CAG	CAG
Glycine	GGC	GGC
		GGT
Histidine	CAC	CAC
Isoleucine	ATC	ATC
Leucine	CTG	CTG
	CTC	
Lysine	AAG	
Phenylalanine	TTC	TTC
Proline	CCC	CCG
Serine	TCC	TCC
	TCG	TCT
	AGC	AGC
Threonine	ACC	ACC
	ACG	ACT
Tyrosine	TAC	TAC
Valine	GTC	
	GTG	

by more than 75% of the organisms within a dataset, then it was marked as preferred for the whole dataset. The preferred codons found for the *E. coli* and Mammalian datasets are given in table 5.3. As has been observed previously it can be seen that most preferred codons have a C or a G in the last position [Nielsen *et al.* 2007]. An analysis of how changes in the cutoff values used affect the results and a comparison to previously published results are given in appendix B.

5.3 Results

In this section the model selection strategies described in section 5.1 are used to determine which of the models implemented in CodonPhyML provide the best fit for each group in the two datasets described above. Due to the large number of possible models and selection strategies not all tests are performed. Besides the Constant Rates model there are 5 different models of rate heterogeneity that can be fitted to each group. Additionally, selection on codon bias can be added to every model. Furthermore, each of these models can be coupled with any of the three frequency models. Lastly, the number of classes can be varied in models with rate heterogeneity. This gives a total of 36 possible models with a fixed number of classes for each group. Exploring different numbers of classes adds another 30 models every time the number of classes are changed. Exploring all of these models is intractable for the amount of groups in the two datasets (542 in the *E. coli* dataset and 912 in the Mammalian dataset).

To simplify matters the Yap frequency model was chosen *a priori*, since it accounts for the background nucleotide content [Yap *et al.* 2010]. Although the Yap model and the Muse-Gaut model give similar results (section 4.2) the Yap model is preferable since the Muse-Gaut model assumes complete independence between the codon positions. Further, the concerns of Yap *et al.* [2010] and Lindsay *et*

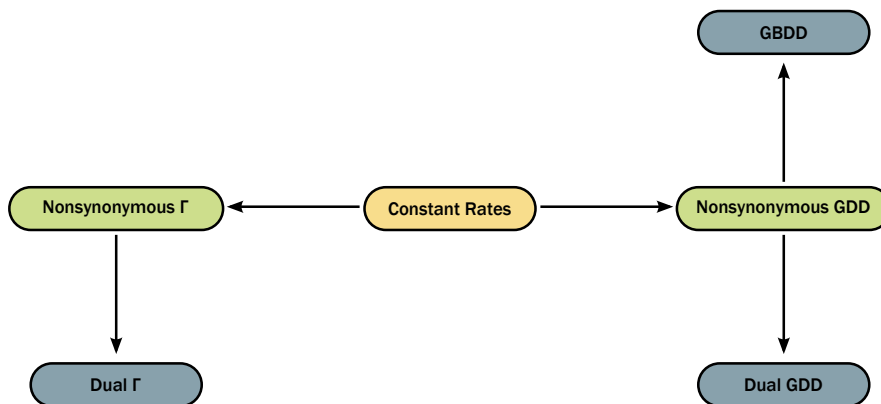


Figure 5.2: The different models of rate heterogeneity that were used on the *E. coli* dataset. Arrows between models indicate a nesting relationship and points toward the more complicated model.

al. [2008] about biases in the Goldman-Yang model may be valid, which makes the Yap model preferable. The model selection procedure discussed in this section was repeated on the Goldman-Yang and Muse-Gaut models. The results are shown in appendix C and indicate that the frequency model does not make a big difference to the model of rate-heterogeneity assigned to a group. Furthermore, it was also found that the frequency model also does not significantly change the rate estimates.

The model selection strategy was further separated into two steps. In the first step of model selection the model of rate heterogeneity is found. This was investigated rigorously on the smaller *E. coli* dataset to reduce the number of models for rate heterogeneity. Only these models were subsequently fit to the Mammalian dataset. This could lead to erroneous results as it is known that there are large differences between prokaryotic and eukaryotic genes. However, due to time constraints, model selection could not be investigated rigorously on the Mammalian dataset as well.

The second step tests for the presence of selection on the codon usage in each of the models found in the first step. This is done by fitting the same model with a ψ parameter added on and using an LRT to determine if the improvement is significant.

All of the experiments below were run 3 times, and only the results from the run with the highest likelihood are shown. In order to save time no topology optimizations were performed. The Goldman-Yang Constant Rates trees described in section 5.2.2 were used as initial trees. To further improve the speed of CodonPhyML, frequencies were not optimized, but counted from the data using $CF3 \times 4$.

The Jaccard index is used to evaluate the similarity between the sets of groups selected by different model selection strategies. The Jaccard index between two sets, A and B , is defined as:

$$\frac{|A \cap B|}{|A \cup B|}$$

Wherever it is reported, the Jaccard index is written as a fraction in order to convey the maximum amount of information to the reader. If the Jaccard index is 1 the sets are identical, if it is 0 the sets are mutually exclusive.

5.3.1 *E. coli* dataset

Investigating rate heterogeneity

As shown in figure 5.2, starting from the Constant Rates model three paths of nested models can be traced to models that allow synonymous variation. The choice was made to select only the two paths resulting in the best overall fit, and to do all further analyses only on the models in these paths. The decision to choose two paths of nested models is motivated by the fact that the use of LRTs for model selection is usually a more rigorous solution than information theoretic criteria. Two paths were chosen to minimize the impact of a particular choice on the LRT procedure. For models with a Γ distribution partitions of the distribution into 3, 6 and 9 classes were investigated. It should be stressed that these models are in effect identical, since they have the same number of free parameters, and only differ

Table 5.4: The results of using information theoretic criteria to select between the 10 candidate models of rate heterogeneity on the *E. coli* dataset.

	Constant Rates	Nonsynonymous GDD	Nonsynonymous Γ			Dual GDD	General Bivariate	Dual Γ		
Number of Rate Parameters	2	6	3			10	8	4		
Number of Classes	1	3	3	6	9	3	3	3	6	9
Assigned by:										
AIC	106	0	0	0	16	4	33	43	89	228
AIC' _c	106	0	0	0	16	4	33	43	89	228
AIC _c	154	0	0	0	16	3	17	32	77	220
BIC	215	0	0	0	14	0	4	17	64	205
High support ($\Delta_i \leq 2$) by:										
AIC	146	1	37	52	52	6	64	181	359	407
AIC' _c	146	1	37	52	52	6	64	181	359	407
AIC _c	183	0	12	34	35	4	29	148	328	377
BIC	235	0	5	20	26	0	4	86	252	308

in the smoothness of the approximation to the Γ distribution. For GDD models only 3 classes were used throughout, since these models do not scale well (figure 4.4) and because these models have more freedom on the distribution used, and can therefore fit more variation with fewer classes. This gives a total of 10 model setups, which were fit to every group in the *E. coli* dataset using the Yap frequency model. CodonPhyML failed to find an optimum on 23 of the groups. These groups were disregarded in all further analyses.

As an exploratory analysis the different information criteria described in section 5.1.2 are used to evaluate model fit. LRTs cannot be used to compare all the different models, since not all of them are nested. The results are summarized in table 5.4. It can be seen that the AIC and AIC'_c give exactly the same results. It was verified that the sets they return are in fact the same. This is not surprising, as the ratio of the sample size to the number of free parameters is never less than 40 for any groups when the sample size is set to the number of codons in the alignment. Thus, the correction made by the AIC'_c is insignificant. On the other hand, when the sample size is set to the number of sites only, the correction should be applied to nearly every group, which is why using the AIC_c makes a big difference to which groups are selected.

Of the three criteria examined here the BIC is the most conservative and assigns more groups to simpler models. The AIC is the least conservative and as expected, the correction applied to the AIC_c leads to it being more conservative. Although the number of groups selected by different criteria varies substantially, the intersection between the groups assigned to a given model by two different

criteria usually contains most of the groups in the smaller set. (This does not hold when one of the sets is very small). It should be kept in mind that the BIC, AIC and AIC_c are all equivalent to the likelihood difference when selecting for the number of classes in the models with a Γ distribution.

From table 5.4 it is clear that the majority of groups prefer either the Constant Rates model or a model with synonymous variation. Most of the groups are assigned to the Dual Γ model, with very few groups being assigned to one of the GDD models. This pattern is repeated when groups with high support are examined, however the Nonsynonymous Γ model with more classes also have non-negligible levels of support. The Dual Γ model with 6 and 9 classes have the highest support among models allowing rate variation. This is probably due to the freedom this model has in letting the rates vary, while still having very few parameters. The Dual Γ model has only 4 rate parameters (regardless of the number of classes) compared to 10 for the Dual GDD and 8 for the GBDD models with 3 classes⁵. Between the GDD models the GBDD model has the highest support. This is due to its ability to capture the same variations in the data as the Dual GDD model while using fewer groups, as reported in [Kosakovsky Pond *et al.* 2010b].

From this exploratory analysis a few tentative conclusions may be drawn. Models that only allow nonsynonymous rate variation have little or no support. Groups either show no rate variation at all, or variation in both the synonymous and nonsynonymous rates. Among GDD models the GBDD model provides the best fit, and among Γ models the Dual models with more classes provide better fits. These conclusions are investigated in more detail below.

The Nonsynonymous Γ and Dual Γ models are examined with respect to the objectives of choosing a model selection strategy and determining how many classes give the best overall fit. Since the Constant Rates, Nonsynonymous Γ and Dual Γ models are nested, if the same number of nonsynonymous classes is used, forward selection and backward elimination are used in addition to information theoretic criteria. Because the AIC'_c is nearly equivalent to the AIC on this dataset, it is not used on any further analyses.

In the first experiment the Nonsynonymous Γ and Dual Γ models are examined separately for different numbers of classes and compared to the fit on the Constant Rates model. The aim of this experiment was to determine which model is preferred between the Constant Rates, Nonsynonymous Γ and Dual Γ models. The three different model selection paths are referred to as Γ_3 , Γ_6 and Γ_9 , depending on the number of classes. The results are summarized in table 5.5. For the information theoretic criteria, the same trends as before are observed. It can also be seen that forward selection is the most conservative selection strategy. On the other hand, the results of backward elimination are very similar to the AIC. This raises some doubts on the validity of backward elimination on this dataset, since for small sample sizes the AIC_c is recommended. From table 5.5 it seems as if varying the number of classes does not have a large effect on the model choice. This is visualized better in heatmaps of the Jaccard indices, shown in figure 5.3. It can be seen that the groups selected by the same selection strategy stays almost stationary as the number of classes are varied. The similarity between the AIC and backward elimination is also clearly visible on the heatmaps.

In a second experiment, the effect of varying the number of classes is investigated by examining the magnitude of the likelihood increases as the number of classes are increased from 3 to 6, 3 to 9 and 6 to 9 on the Dual Γ model. This was not done for the Nonsynonymous Γ model as not enough groups are assigned to it for the results to be meaningful. It should be noted that in this case the likelihood increase is equal to any of the information criteria as the models all have the same number of free parameters. The only difference between the models is that a better approximation to the Γ distribution is used (equation 3.10). The results are shown in figure 5.4. Assuming that a likelihood increase of less than 2 is not significant (table 5.1) it can be concluded that increasing the number of classes from 3 to 6 and 3 to 9 classes leads to a better fit in about half the groups. However, increasing the number of classes from 6 to 9 only results in a significant increase in the likelihood in a small percentage of the groups. Hence, it can be concluded that 6 classes give the best fit for a Dual Γ model.

The two model selection paths with GDD models are compared to each other and the different model selection strategies. The two paths are referred to as the Dual GDD and GBDD paths. The

⁵ κ is included as one of the rate parameters, in addition to the parameters listed in figures 4.1 and 4.2.

Table 5.5: Results of the different selection strategies on the *E. coli* dataset for models where rate heterogeneity is treated with a Γ distribution, for different numbers of classes.

Selection Strategy	Constant Rates				Nonsynonymous variation				Nonsynonymous and Synonymous variation						
	Groups Assigned (of 519)	High Support ($\Delta_i \leq 2$)	Jaccard Index wrt AIC	Jaccard Index wrt AIC _c	Jaccard Index wrt BIC	Groups Assigned (of 519)	High Support ($\Delta_i \leq 2$)	Jaccard Index wrt AIC	Jaccard Index wrt AIC _c	Jaccard Index wrt BIC	Groups Assigned (of 519)	High Support ($\Delta_i \leq 2$)	Jaccard Index wrt AIC	Jaccard Index wrt AIC _c	Jaccard Index wrt BIC
Γ_3 path															
AIC	120	163				13	54				386	430			
AIC _c	170	195				10	31				339	385			
BIC	237	261				10	22				272	302			
Forward Selection*	393		120/393	169/394	231/399	9		7/15	7/12	6/13	117		117/386	117/339	110/279
Backward Elimination*	126		114/132	123/173	126/237	9		7/15	7/12	6/13	384		378/392	338/385	272/384
Γ_6 path															
AIC	108	156				16	54				395	435			
AIC _c	161	186				12	38				346	397			
BIC	218	240				14	22				287	316			
Forward Selection*	382		108/382	161/382	216/384	10		8/18	9/13	8/16	127		127/395	127/346	121/293
Backward Elimination*	117		102/123	113/165	117/218	10		8/18	9/13	8/16	392		386/401	344/394	287/392
Γ_9 path															
AIC	107	149				17	53				395	435			
AIC _c	156	183				17	37				346	397			
BIC	216	236				14	26				289	318			
Forward Selection*	360		107/360	156/360	208/368	15		13/19	14/18	8/21	144		144/395	143/347	137/296
Backward Elimination*	114		104/117	111/159	114/216	15		13/19	14/18	8/21	390		387/398	344/392	289/390

* All tests performed at $\alpha = 0.05$

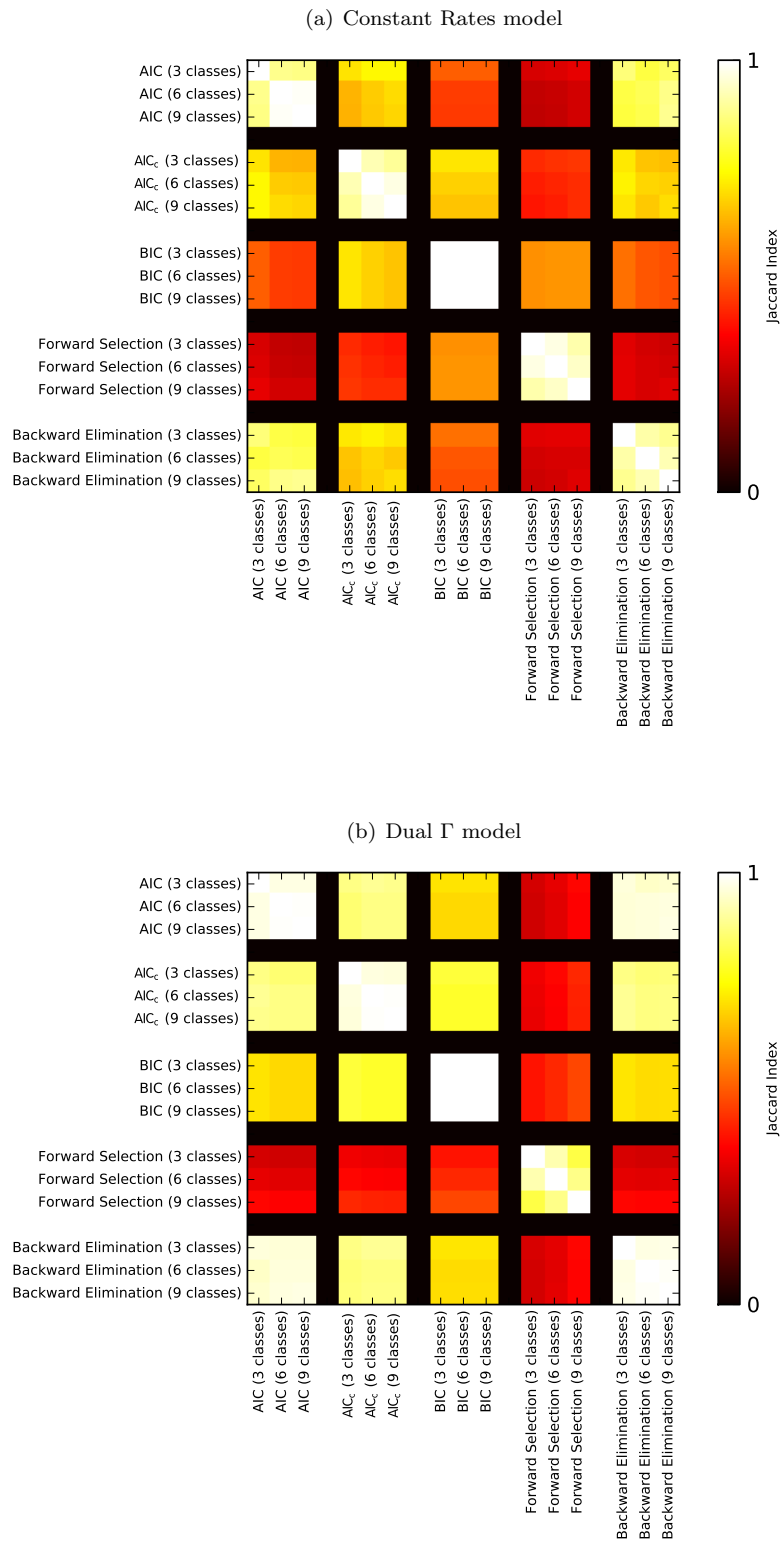


Figure 5.3: Heatmap of the Jaccard indices between the sets of groups selected by different strategies for the Constant Rates and the Dual Γ model on the *E. coli* dataset. Results for the Nonsynonymous Γ model are not shown as not enough groups were assigned to this model for the results to be meaningful.

Table 5.6: Results of the different selection strategies on the *E. coli* dataset for models where rate heterogeneity is treated with a GDD distribution.

Selection Strategy	Constant Rates				Nonsynonymous variation				Nonsynonymous and Synonymous variation						
	Groups Assigned (of 519)	High Support ($\Delta_i \leq 2$)	Jaccard Index wrt AIC	Jaccard Index wrt AIC _c	Jaccard Index wrt BIC	Groups Assigned (of 519)	High Support ($\Delta_i \leq 2$)	Jaccard Index wrt AIC	Jaccard Index wrt AIC _c	Jaccard Index wrt BIC	Groups Assigned (of 519)	High Support ($\Delta_i \leq 2$)	Jaccard Index wrt AIC	Jaccard Index wrt AIC _c	Jaccard Index wrt BIC
Dual GDD path															
AIC	246	267				17	30				256	272			
AIC _c	300	313				13	20				206	221			
BIC	409	420				11	14				99	106			
Forward Selection*	411		246/411	295/416	372/448	12		11/18	10/15	2/21	96		96/256	91/211	60/135
Backward Elimination*	206		199/253	206/300	206/409	12		11/18	10/15	2/21	301		254/303	206/301	99/301
GBDD path															
AIC	199	234				10	14				310	328			
AIC _c	257	270				6	8				256	276			
BIC	362	372				3	5				154	164			
Forward Selection*	411		199/411	255/413	340/433	7		6/11	6/7	1/9	101		101/310	100/257	83/172
Backward Elimination*	141		137/203	141/257	141/362	7		6/11	6/7	1/9	371		309/372	256/371	154/371

* All tests performed at $\alpha = 0.05$

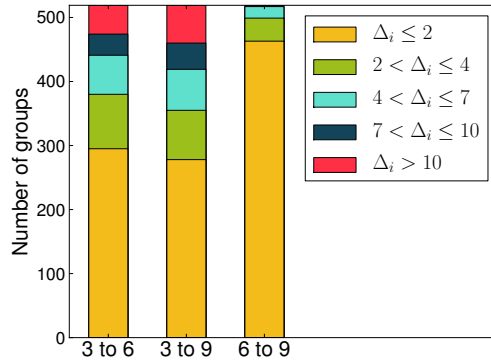


Figure 5.4: Histograms of the magnitude of the likelihood increase as the number of classes are increased on the Dual Γ model.

results are given in table 5.6 and are similar to those found for the Γ models. However, there are some small differences. For GDD models most selection strategies are more conservative than with Γ models, resulting in fewer groups being assigned to models with rate variation. Another difference is that the AIC is more conservative than backward elimination here. This observation adds strength to the assumption that backward elimination chooses models that are too complicated. When the heatmaps (figure 5.5) are examined it is seen that the sets of models assigned along the Dual GDD path and the GBDD path are similar. However, when comparing the Dual GDD and GBDD models to each other using the information theoretic criteria the GBDD model is preferred for more than 500 groups regardless of which criterion is used. In fact, the Dual GDD model never receives support in more than 50 groups, even when the least conservative criterion (AIC) is used. This backs up the claim by Kosakovsky Pond *et al.* [2010b] that the GBDD provides a better fit than the Dual GDD model, due to it having as much freedom in choosing rates as the Dual GDD model, but fewer rate classes.

Based on the above results the Γ_6 and GBDD model selection paths are chosen. The decision for the Γ_6 path was motivated by the fact that there is no overall improvement when increasing the number of classes from 6 to 9, as there is for increasing the classes from 3 to 6. The GBDD path was chosen as it clearly gives a better fit than the Dual GDD path. In figure 5.7 it can be seen that the sets of groups assigned to the Constant Rates model and models allowing synonymous variation are similar for the two selection paths. Since both paths only assign a small number of groups to models that only allow nonsynonymous variation, these sets are not expected to show a high degree of similarity.

One drawback of the Dual Γ model is that the Γ distribution is approximated by classes with equal weights. This is not what is expected, and is also not what is seen with the GBDD model. For the Dual Γ model with 6 classes, in each group, a sixth of the sites will be assigned to each synonymous rate. However, in the GBDD model with 3 classes, in almost every group one of the synonymous rate classes receive a weight of less than a tenth. Thus, it is necessary to use more classes to get an equivalent description of the data when using a Γ model instead of a GDD model. The histograms of the nonsynonymous and synonymous coefficients of variation for both the Dual Γ and GBDD models are shown in figure 5.6. It can be seen that the synonymous coefficient of variation is concentrated between 1 and 2 for both models, showing that most of these groups do indeed have a fair amount of synonymous variation. The nonsynonymous coefficient of variation differs quite a lot between the models. The GBDD model sporadically finds very large values for it, which probably points to one or more of the detected rates having diverged. Hence, it can be concluded that the Γ models are more stable and faster to fit, however GDD models have a more flexible distribution, which can fit more variation. Unfortunately, this also results in them being more unstable.

It appears that backward elimination assigns groups to models that are too complicated, hence forward selection is preferred. Forward selection is also preferred above the information theoretic criteria as it gives a sound selection strategy along a path of nested models. Backward elimination is most effective when it is used to eliminate non-significant parameters one by one from a set of candidate

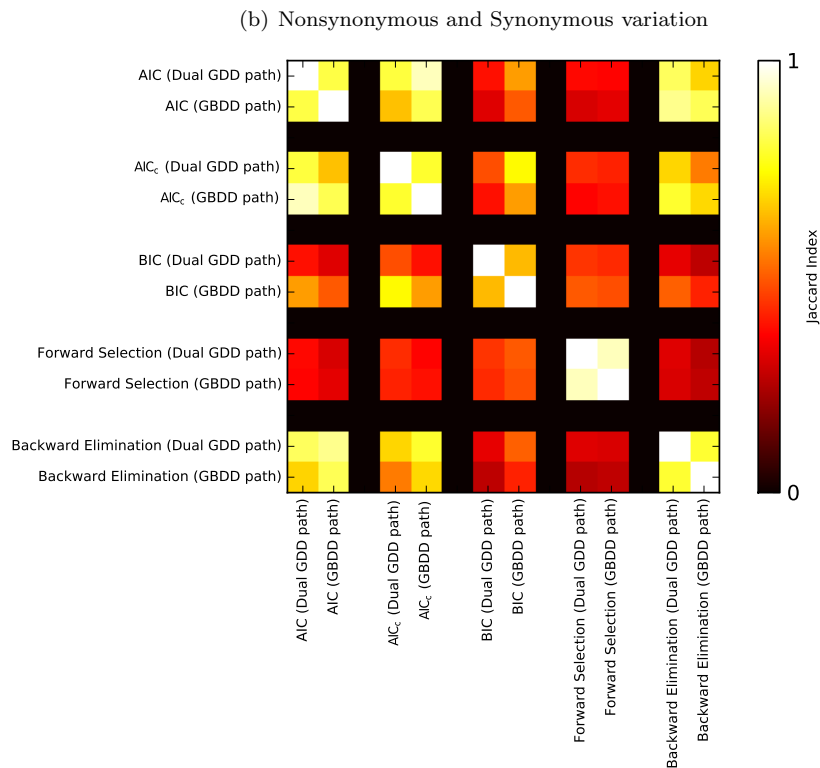
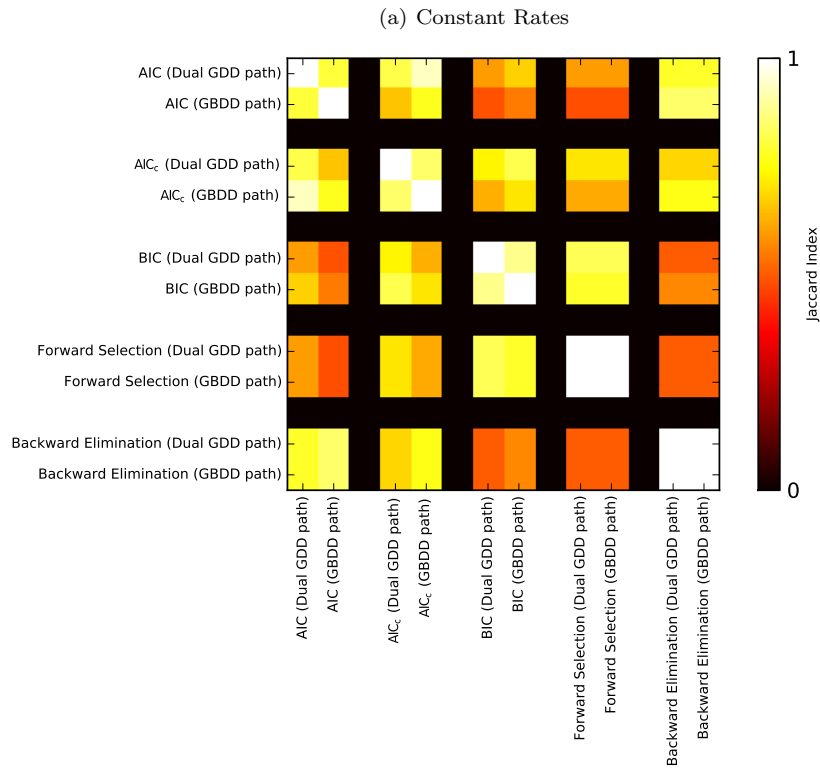
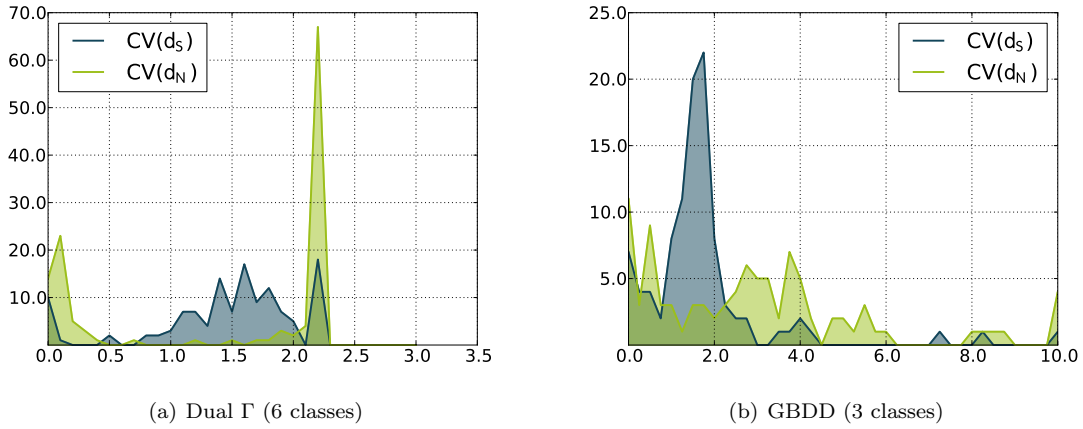


Figure 5.5: Heatmap of the Jaccard indices between the sets of groups selected by different strategies for the Constant Rates model and models that allow synonymous variation along a GDD on the *E. coli* dataset. Results for the Nonsynonymous GDD model are not shown as not enough groups were assigned to this model for the results to be meaningful.

E. coli dataset



Mammalian dataset

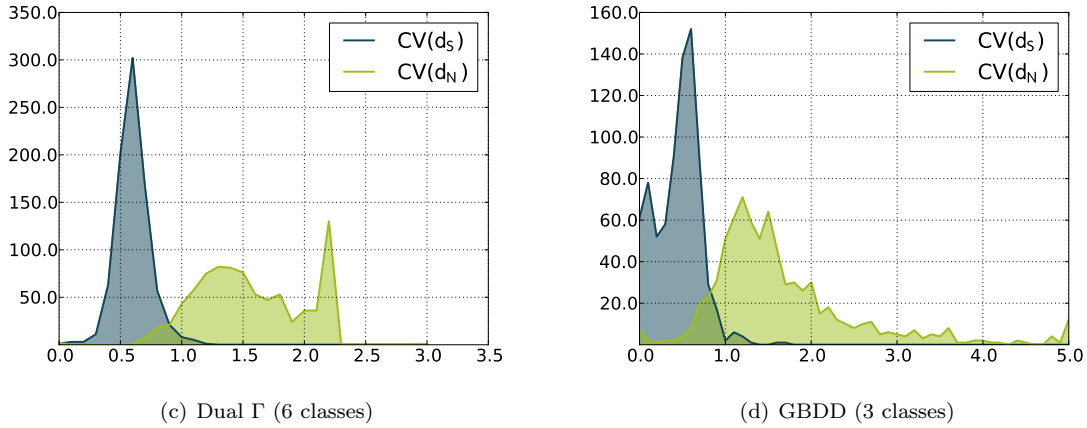


Figure 5.6: Histograms of the synonymous and nonsynonymous coefficients of variation on the groups selected for the Dual Γ and GBDD models.

models. The nested models examined here differ in several parameters, which makes it difficult to identify exactly which parameters are superfluous. Backward elimination will prefer parameter-rich models when there is a large gap between the number of parameters, as is the case between the Nonsynonymous GDD and Dual GDD or GBDD models. The results from forward selection are also more in line with what has been found previously, assigning roughly a fifth of the groups to models with synonymous rate variation [Dimitrieva and Anisimova 2011]. In accordance with these observations Bao *et al.* [2007] reported that both backward elimination and the AIC_c and the AIC_c are biased toward more complicated models.

Selection on codon usage

An LRT was employed to find genes where selection on the codon usage plays a significant role. The test was performed on each of the models in the two model selection paths chosen in the previous section. The numbers of groups selected for each of the five models when the ψ parameter is introduced are shown in table 5.7. It can be seen that the number of groups selected stays relatively constant for the different models, although the set shrinks as the base model gets more complicated. As is expected, overlaps between the sets of selected groups are very large (figure 5.7).

Figure 5.8a shows the distribution of ψ on the Constant Rates model. It was further found that the distribution of ψ values found does not change significantly for different models of rate heterogeneity.

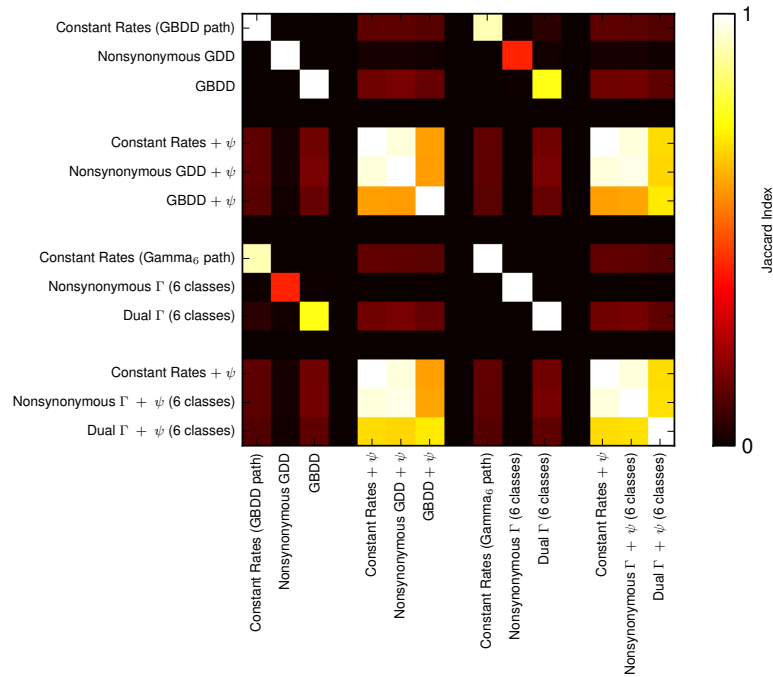


Figure 5.7: Heatmap of the Jaccard indices between the sets of groups that show selection on the codon usage and the sets of groups selected by forward selection for various models of rate heterogeneity on the *E. coli* dataset.

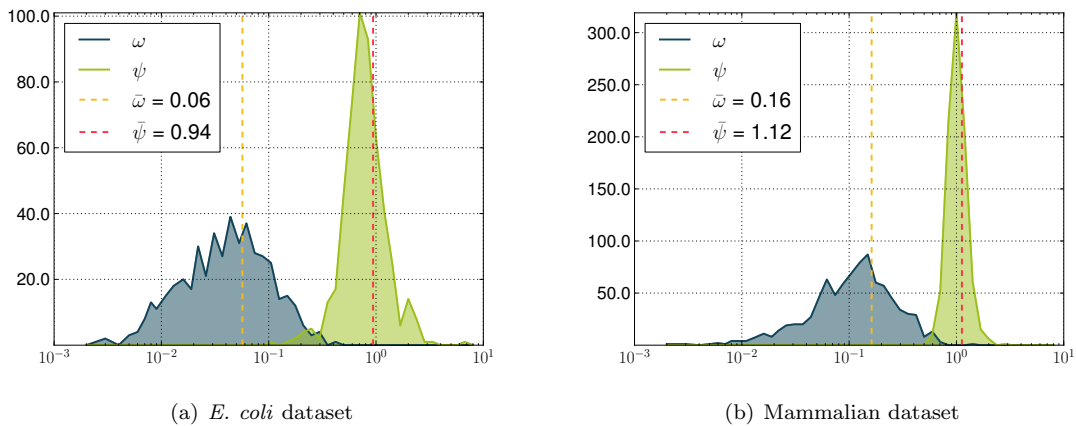


Figure 5.8: Histograms of ω and ψ values estimated with the Constant Rates + ψ model. Note that the x -axis is on a logarithmic scale. It can be observed that the Mammalian dataset has a much higher mean ω rate. This is probably due to the larger sequence divergence. Further it can be observed that whereas the codon usage is on average under purifying selection in the *E. coli* dataset, it is on average under diversifying selection in the Mammalian dataset.

The distribution is centered around 0.94, with fairly fat tails, and some weight above 1, indicating that there is some evidence of diversifying selection on the codon usage. Additionally, it was found that the distribution of ψ in groups where the improvement is significant does not differ much from the distribution over all groups, although the values are slightly lower, indicating stronger purifying selection on the codon usage in these groups. The OMA group descriptions of groups where $\psi > 1.25$ or $\psi < 0.75$ are listed in appendix D.

Figure 5.8a also shows the distribution of ω , which is generally much lower than ψ . This parallels the findings of Zhou *et al.* [2010], who concluded that nonsynonymous substitutions accumulate approximately an order of magnitude slower than nonconservative synonymous substitutions. Additionally, the distributions for ω and κ remain unchanged with the addition of ψ , which shows that the parameters are not being confounded. To verify that ψ is not confounded with any of the nonsynonymous or synonymous rate parameters the overlap between the groups selected here and groups selected for the models with rate heterogeneity was examined. Figure 5.7 shows that there are no correlations between the sets.

5.3.2 Mammalian dataset

Investigating rate heterogeneity

Rate heterogeneity is only investigated on the two model selection paths identified for the *E. coli* dataset, which gives a total of 5 models. CodonPhyML failed to find an optimum in 28 of the groups, which were subsequently disregarded. On the Mammalian dataset $\frac{n}{k} < 40$ on most groups when n is set equal to the number of sites in the alignment, indicating again that the AIC_c would be a better criterion than the AIC. As with the *E. coli* dataset this condition again almost never holds when n is set equal to the number of codons in the alignment, which means that using the AIC'_c would give very similar results to using the AIC.

The results of using different selection strategies on the two model selection paths are shown in table 5.8. The biggest difference is in the amount of groups that show evidence of synonymous rate variation. All of the strategies employed find synonymous variation in more than three quarters of the groups. The Constant Rates model and models allowing only nonsynonymous variation both receive similar levels of support. Furthermore, in accordance with the results on the *E. coli* dataset, the GBDD path is more conservative than the Γ_6 path. While a large overlap can be seen between groups assigned to the GBDD and Dual Γ models, the overlap for other pairs of models is relatively small (figure 5.9). This is probably caused by the small number of groups assigned to these models.

Relative to each other the model selection strategies perform for the most part in a similar manner to what was observed on the *E. coli* dataset. However, the differences between strategies are smaller. In particular, the difference between the AIC and the AIC_c is practically non-existent. Furthermore, the BIC is more conservative than forward selection on the Mammalian dataset.

On the Mammalian dataset the synonymous coefficient of variation is on average significantly lower than on the *E. coli* dataset (figure 5.6). It can be concluded that although more groups with syn-

Table 5.7: The number of groups in the *E. coli* dataset showing significant improvement when the ψ parameter is added to different models.

Model	Significant groups (of 519)*
Constant Rates + ψ	82 (15.80%)
Nonsynonymous GDD + ψ (3 classes)	83 (15.99%)
Nonsynonymous Γ + ψ (6 classes)	81 (15.61%)
GBDD + ψ (3 classes)	72 (13.87%)
Dual Γ + ψ (6 classes)	67 (12.91%)

* All tests performed at $\alpha = 0.05$

Table 5.8: Results of the different selection strategies on the Mammalian dataset for the two model selection paths that were used to model rate heterogeneity.

Selection Strategy	Constant Rates				Nonsynonymous variation				Nonsynonymous and Synonymous variation					
	Groups Assigned (of 884)	High Support ($\Delta_i \leq 2$)	Jaccard Index wrt AIC	Jaccard Index wrt AIC _c	Jaccard Index wrt BIC	Groups Assigned (of 884)	High Support ($\Delta_i \leq 2$)	Jaccard Index wrt AIC	Jaccard Index wrt AIC _c	Groups Assigned (of 884)	High Support ($\Delta_i \leq 2$)	Jaccard Index wrt AIC	Jaccard Index wrt AIC _c	Jaccard Index wrt BIC
Γ_6 path														
AIC	6	6				12	22				866	879		
AIC _c	6	6				14	27				864	871		
BIC	6	9				30	45				848	862		
Forward Selection*	22		6/22	6/22	6/22	19		12/19	13/20	19/30	843/866	842/865	832/859	
Backward Elimination*	3		3/6	3/6	3/6	19		12/19	13/20	19/30	859/869	858/868	848/862	
GBDD path														
AIC	12	13				53	68				819	838		
AIC _c	19	23				60	83				805	820		
BIC	33	37				129	154				722	751		
Forward Selection*	39		12/39	17/41	27/45	66		52/67	56/70	64/131	779/819	774/810	710/791	
Backward Elimination*	11		10/13	11/19	11/33	66		52/67	56/70	64/131	805/821	796/816	722/807	

* All tests performed at $\alpha = 0.05$

Table 5.9: The number of groups in the Mammalian dataset showing significant improvement when the ψ parameter is added for different models.

Model	Significant groups (of 884)*
Constant Rates + ψ	233 (26.36%)
Nonsynonymous GDD + ψ (3 classes)	265 (29.98%)
Nonsynonymous Γ + ψ (6 classes)	245 (27.71%)
GBDD + ψ (3 classes)	384 (43.44%)
Dual Γ + ψ (6 classes)	315 (35.63%)

* All tests performed at $\alpha = 0.05$

onymous variation are found, the variation in these groups is much less pronounced than in *E. coli* data. This probably indicates that there are only a few sites that are under selection. In contrast to the *E. Coli* dataset, the nonsynonymous coefficient of variation has a more peaked distribution for Mammalian data, although the distribution is still flatter than the synonymous coefficient of variation. The reason for this behaviour is probably due to a combination of two factors. Firstly, the nonsynonymous coefficient of variation is more variable than the synonymous coefficient of variation. Secondly, the sets in the Mammalian dataset are about 7 times larger. It is probable that if more groups with synonymous variation were found for the *E. coli* dataset, its nonsynonymous coefficient of variation would show a similar distribution, but shifted to the right. It can also be concluded from figure 5.6 that variations in the nonsynonymous rate are in general higher than variations in the synonymous rate, indicating that selection plays a stronger role on nonsynonymous sites.

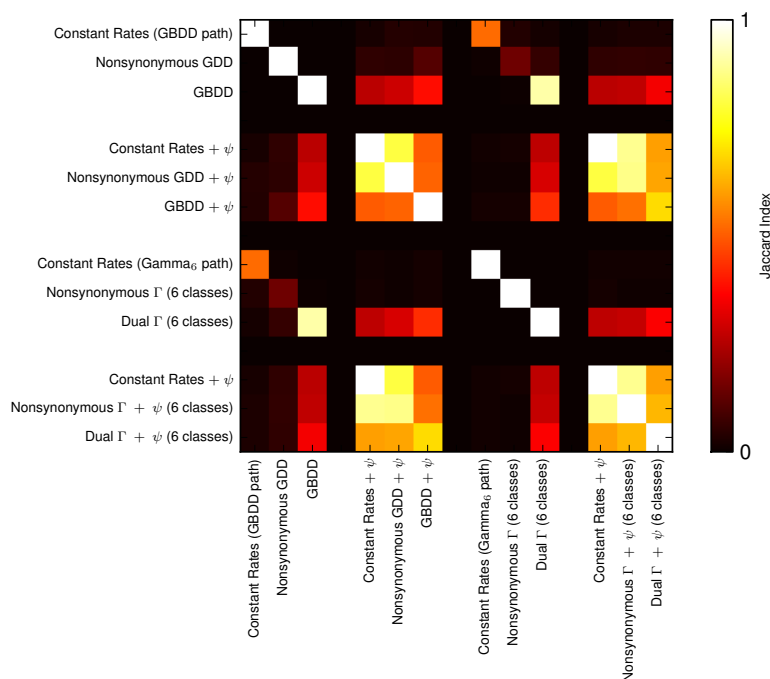


Figure 5.9: Heatmap of the Jaccard indices between the sets of groups that show selection on the codon usage and the sets of groups selected by forward selection on various models of rate heterogeneity on the Mammalian dataset.

Selection on codon usage

The number of groups where selection on codon bias leads to a significant improvement is shown in table 5.9. Proportionally more groups are found than for the *E. coli* dataset. Contrary to what was found for the *E. coli* dataset, the number of groups where adding a ψ parameter led to a significant increase in the likelihood grows with the model complexity. Another difference is that the estimates of ψ cluster around 1.12, indicating more positive selection on the codon usage (figure 5.8b). Nevertheless, there are also a number of groups under purifying selection on the codon usage, and as before groups where $\psi > 1.25$ and $\psi < 0.75$ are listed in appendix D. Although the distribution of ψ is shifted in the Mammalian dataset the distribution of ω is similar to the distribution of ω in the *E. coli* dataset, although the mean rate is also higher. Further, it was also found that for the Mammalian dataset the distribution of ψ again did not significantly change for different models of rate heterogeneity, or when only groups where the improvement was significant were considered. In all cases the distribution remained centered around 1. Lastly, figure 5.9 shows that there is a large overlap between the groups found to have evidence for selection on codon usage with the Constant Rates, Nonsynonymous GDD and Nonsynonymous Γ models. The GBDD and Dual Γ models were both chosen for many more groups, and hence do not show as big an overlap. Figure 5.9 further shows that there is again no evidence that ψ is being confounded with the other rate parameters.

5.4 Summary

Two model selection paths were chosen, with each path containing 5 models. The final assignments of groups to models with forward selection are given in fig 5.4. Based on the results above, the following conclusions may be drawn. Nonsynonymous rate variation almost never occurs by itself. When there is rate variation it is likely that both rates vary. There are more groups with rate variation on the

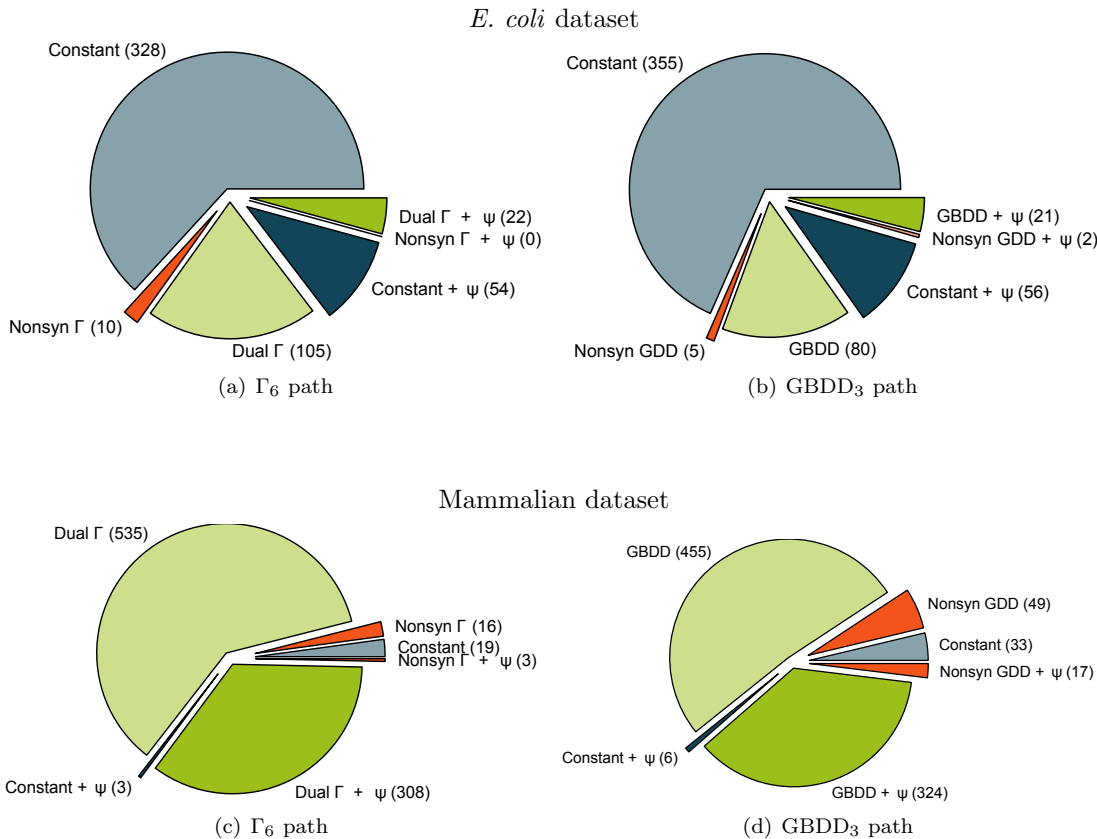


Figure 5.10: The number of groups assigned to each model in the two selection paths that were chosen using forward selection.

Mammalian dataset, but the rate variations are less intense. The higher incidence of rate variation in the Mammalian dataset may be a result of the higher divergence of sequences in the Mammalian dataset. Furthermore, there is evidence for both purifying and diversifying selection on the codon usage in both datasets in a non-negligible amount of groups. Finally, it has also been shown that ψ is not confounded with any of the other rates, but varies by itself.

Chapter 6

Relationships between evolutionary rate variation, protein structure and codon usage

In the previous chapter, an *E. coli* and a Mammalian dataset were divided, based on model selection, into groups with evidence of synonymous rate variation and groups with evidence of selection on the codon usage. In this chapter, a search is carried out for pairwise correlations between the site-specific properties of the sequences from the model organisms of each of the datasets. The following site specific properties are available:

- Amino acid
- Codon
- Secondary structural class
- Nonsynonymous rate
- Synonymous rate

The amino acid frequency, absolute codon frequency, synonymous codon frequency and RSCU can all be calculated trivially from the datasets. Secondary structure assignments are not available for every site, although some assignments are available for every sequence. Sites with a secondary structure assignment are classed into helices, sheets or loops. Although estimates for the nonsynonymous and synonymous rates are available for every site, they should only be used on groups where evidence of rate heterogeneity was detected. For each rate the assigned rate class, the maximum likelihood estimate of the rate and the mean posterior rate are available. Sites were assigned to the rate class with the highest posterior probability at each site, using NEB.

According to the model selection strategy followed in the previous chapter the two datasets were divided into 7 partially overlapping datasets. The datasets are shown in table 6.1 and are named by the model of synonymous rate variation that was used. (Note that (+ ψ) is used to indicate the set of all groups assigned to a model with and without the ψ parameter). Each of these 7 datasets can further be restricted to only sites with secondary structure assignments. There is a large overlap between some of the datasets, for instance GBDD and Dual Γ . This is intentional and can be used to assess how differences between the datasets affect inferences.

6.1 Relationships to structure

Using only sites with structure assignments, χ^2 -tests were employed to test for the differential usage of amino acids and synonymous codons within different secondary structural classes. Additionally, the posterior nonsynonymous and synonymous rates were partitioned into quartiles and the hypothesis that different secondary structures have different mean rates was tested.

Table 6.1: The numbers of groups assigned to site-specific datasets used to search for pairwise correlations in this chapter.

	<i>E. coli K12</i>			<i>Homo sapiens</i>		
	OMA Groups	Total number of sites	Sites with structure assignments	OMA Groups	Total number of sites	Sites with structure assignments
Complete dataset	542	179,392	161,549	912	560,506	234,951
GBDD	80	36,741	31,868	455	251,493	116,258
GBDD + ψ	21	10,510	8,721	324	248,604	92,123
GBDD (+ ψ)	101	47,251	40,589	779	500,097	208,381
Dual Γ	105	44,669	39,877	535	292,435	137,043
Dual Γ + ψ	22	12,024	10,068	308	235,739	84,507
Dual Γ (+ ψ)	127	56,693	49,945	843	528,174	221,550

Testing the preferential usage of any of these quantities is conceptually similar and involves the distribution of n discrete classes into the three secondary structure categories. For the amino acid usage $n = 20$. For synonymous codons n is equal to the number of codons in the synonymous codon family. For nonsynonymous and synonymous rates $n = 4$, because the rates were partitioned into quartiles.

The null hypothesis in tests with the amino acid usage and the evolutionary rates is that the n categories are randomly used in the different secondary structures. Hence, calculating expected frequencies is straightforward. When calculating the expected frequency of a synonymous codon in a secondary structural class the usage of the amino acid it codes for within the structural class and the inherent synonymous codon bias in the sequence need to be taken into account. To do this the method of Adzhubei *et al.* [1996], described in section 2.3.3, was used.

To perform a test on a dataset the following steps were followed. For each of the n classes the expected usages in different secondary structure classes were calculated for each sequence. The expected and observed usages of all the sequences in the datasets were then added, giving an $n \times 3$ contingency table. A χ^2 -test was then used to assess if there is an overall effect present between the quantity of interest and the secondary structural classes. All of the tests described in the following sections were performed at a significance level of $\alpha = 0.05$.

6.1.1 Relationship between amino acid usage, codon usage and secondary structure

A significant effect was detected in all of the datasets for both *E. coli K12* and *Homo sapiens* with respect to the amino acid usage. However, a correlation between the amino acid usage and the secondary structure is not interesting, but expected. This is a well-known effect and it has been shown that the bias between amino acid usage and the secondary structure is highly correlated to the physico-chemical properties of amino acids [Chiusano *et al.* 2000, Chou and Fasman 1974]. Hence, these correlations were not investigated further to find which sequences have the largest biases.

On the other hand, none of the synonymous codon families showed a significant deviation from the expected usage on any of the datasets for either of the organisms. In order to look for a weaker effect

Table 6.2: Codons that show a significant deviation from their expected usages within their synonymous codon families. The datasets are as defined in table 6.1.

	<i>E. coli K12</i>	<i>Homo sapiens</i>
Complete dataset	GCG, GCT, GAG, GGG, CTC, CTG, TTA, ACA	GGC, GGG, TCC
GBDD	CTC, CTG, TCC, GTC	TCC, TCG
GBDD + ψ		GGG
GBDD (+ ψ)	GAG, CTC, CTG, TCC, GTC	GGG, TCC
Dual Γ	CTC, CTG, TCC	TTA, CCC, TCC
Dual Γ + ψ	GGG, TCA	CGC
Dual Γ (+ ψ)	GGG, CTG, TCC	GGC, GGG, TCC

an independent test was performed on each synonymous codon. In this manner a significant effect was detected on several codons, for both organisms. The codons that tested significant are listed in table 6.2.

6.1.2 Relationship between evolutionary rates and secondary structure

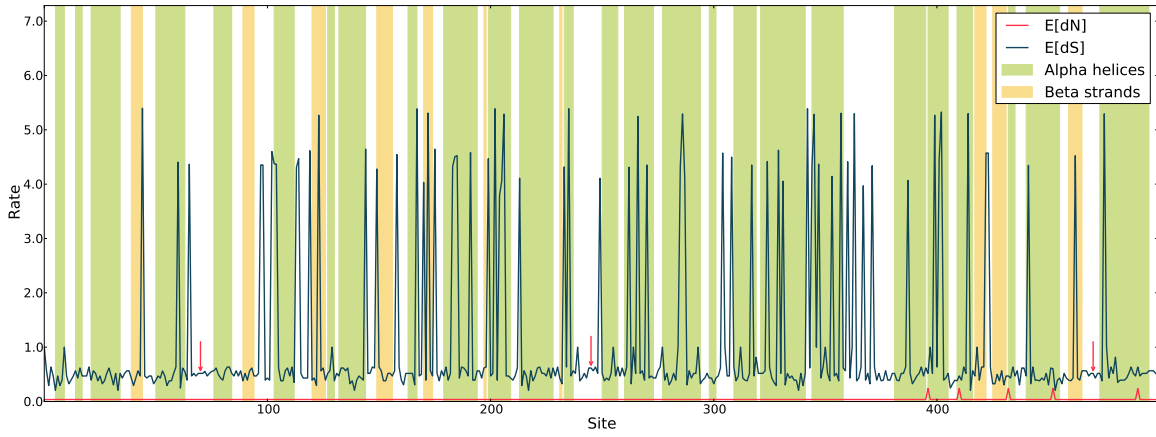
Tests for an effect between the posterior evolutionary rates and the secondary structure gave significant results in most cases. Unlike the tests involving the amino acid and codon usages these tests were only performed on datasets with evidence of synonymous variation. The only negative result was for the test between the nonsynonymous rate and the secondary structure on the Dual Γ (+ ψ) dataset for *E. coli K12*. All of the other tests reported a significant bias, in both the nonsynonymous and synonymous rates, on all the other datasets, for both organisms.

This is evidence of the existence of a relationship between the evolutionary rates and the secondary structure. To determine which sequences have the strongest bias the tests were repeated on each sequence individually. To ascertain that the results are statistically significant, all sequences where any of the expected values in the contingency table are below 5 were deemed inadmissible. Despite this restriction, a relatively large number of sequences show a significant bias (table 6.3). The 10 sequences with the highest χ^2 -test statistics for the GBDD (+ ψ) and Dual Γ (+ ψ) models for both organisms are listed in appendices E and F.

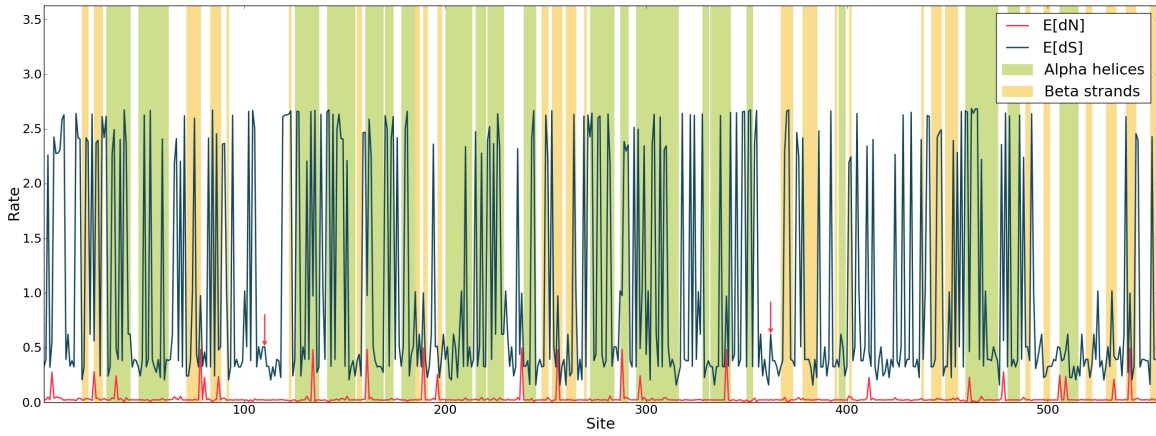
It can be seen from table 6.3 that biases occur more frequently for the nonsynonymous rate. This is to be expected, as nonsynonymous mutations are more likely to destroy secondary structures. Hence, nonsynonymous substitutions within secondary structures should be selected against. On the other hand, biases in the synonymous rate may point to purifying selection acting on the elongation rate, or selection for preferred codons. The effects detected for the nonsynonymous and synonymous rates are unrelated to each other as the number of groups showing biases for both rates is never particularly high.

An examination of the rate profiles indicates that the real situation is more complex than the simple explanation offered above. In most cases a clear effect is not visible. Especially for the nonsynonymous rate most of the significant sequences seem to have only a few sites with high nonsynonymous rates, with the rest of the sites being effectively invariable. In such cases the effect that was detected is most probably an artefact of partitioning the rates into quartiles. Nevertheless, weak effects may be observed in some sequences. In figure 6.1a and b an effect was detected with respect to the synonymous rate. Figure 6.1a appears to have a low synonymous rate in the regions between secondary structures around site 70, 245 and 470. Similarly, in figure 6.1b, there is a long gap between two beta strands around site 110 with a significantly long stretch with low synonymous rates. Another, smaller gap with

(a) *E. coli* K12 DNA polymerase III (OMA group 34912) (Dual Γ (+ ψ) dataset)



(b) *E. coli* K12 Gamma-glutamyltranspeptidase (OMA group 35724) (GBDD (+ ψ) dataset)



(c) *Homo sapiens* Heat shock 40 kDa protein 1 (OMA group 76612) (GBDD (+ ψ) dataset)

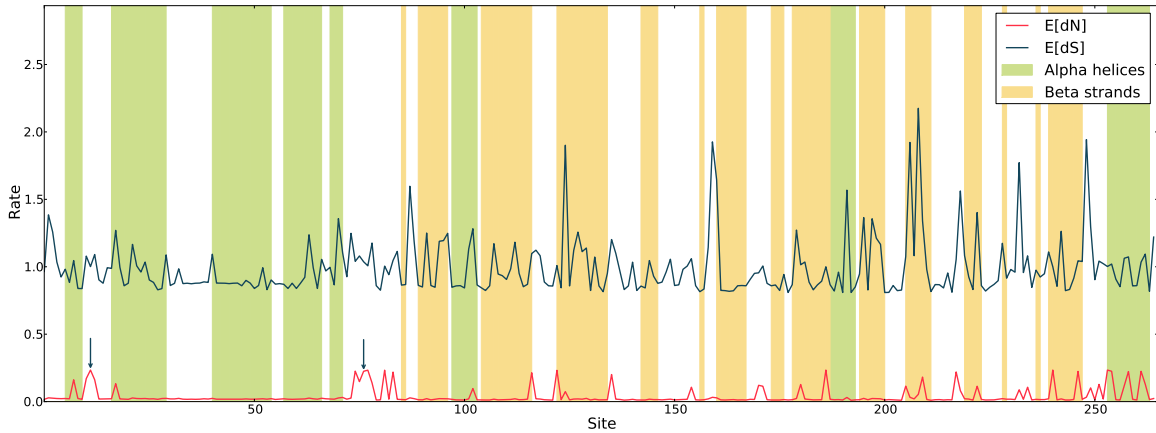


Figure 6.1: Plots of the posterior nonsynonymous and synonymous rates with respect to secondary structures. In (a) and (b) an effect with respect to the synonymous rate was detected. In (c) an effect was detected with respect to the nonsynonymous rate.

Table 6.3: The number of sequences that show a significant bias between the posterior evolutionary rates and the secondary structure. The datasets are as defined in table 6.1.

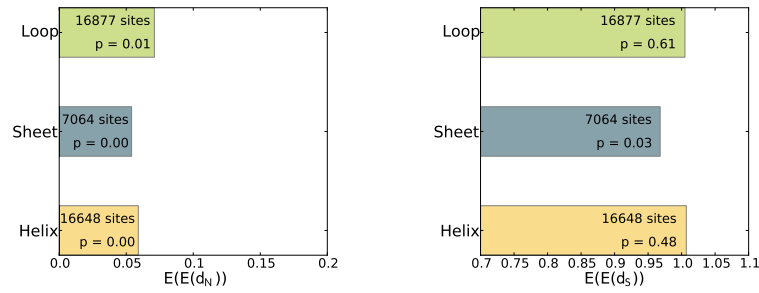
	<i>E. coli K12</i>		<i>Homo sapiens</i>	
	$\mathbb{E}(d_N)$	$\mathbb{E}(d_S)$	$\mathbb{E}(d_N)$	$\mathbb{E}(d_S)$
GBDD	19	11	56	38
GBDD + ψ	6	4	38	23
GBDD (+ ψ)	25	15	94	61
Dual Γ	49	20	108	28
Dual Γ + ψ	7	7	52	6
Dual Γ (+ ψ)	56	27	160	34

a low synonymous rate is observable around site 362. Stretches with a low synonymous rate between secondary structures indicates purifying selection for the synonymous codon choice. These sites may indicate the presence of translational pauses. This could be investigated by examining the codon usages and mRNA secondary structures at these sites. In both these figures it can also be seen that most sites with a high synonymous rate occur within secondary structures, which indicates diversifying selection in these regions. In figure 6.1c it is clear that most sites with a high nonsynonymous rate fall between secondary structural elements. In particular, there are stretches around sites 10 and 75 that seem to be under diversifying selection.

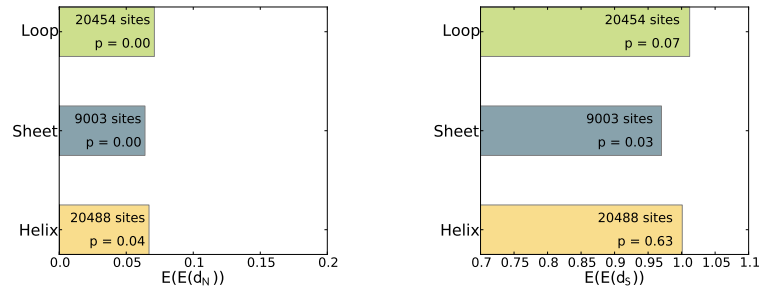
From the figures it can also be seen that the synonymous rate is much higher than the nonsynonymous rate most of the time. This is to be expected, since nonsynonymous mutations affect the structure and function of a protein, hence most nonsynonymous mutations are deleterious [Smith and Simmonds 1997]. If the mutation is not viable, the organism will not survive and there will be no evidence of the mutation left. Hence, the nonsynonymous rate is expected to be much lower, because successful nonsynonymous mutations are much rarer.

In table 6.1.2 the means of the mean posterior rates are shown for the different secondary structural elements for the GBDD (+ ψ) and Dual Γ (+ ψ) datasets. The results for all the datasets are given in appendix G. The significance of the deviations of the mean posterior rates on the different secondary structural elements from the mean posterior rates on the complete datasets was assessed with a Z -test. It can be seen that the nonsynonymous rate is on average higher for loops on all of the datasets for both organisms, as postulated above. The deviation is significant for all of the datasets except GBDD + ψ for *E. coli K12*. However, this is also the smallest of all the datasets (table 6.1). It can be concluded that loops are indeed under less pressure from purifying selection. It seems as if sheets are under stronger purifying selection than helices for *E. coli K12*, however this does not seem to be the case for *Homo sapiens*. The reason may be because sheets rely on longer distance interactions to form, and hence may not be as tolerant to mutations as helices. Regarding the synonymous rate in *E. coli K12* only sheets show a significant deviation on more than one dataset and seem to be under pressure to accept only a few synonymous mutations. It should be kept in mind that fewer sites are assigned to sheets than the other secondary structures, hence deviations are not as significant as for the other classes. In *Homo sapiens* the deviations are more significant and again point to purifying selection on sheets. Furthermore, in *Homo sapiens*, only loops have an average synonymous rate greater than 1, indicating that the codon usage in secondary structures is more stationary than in loops.

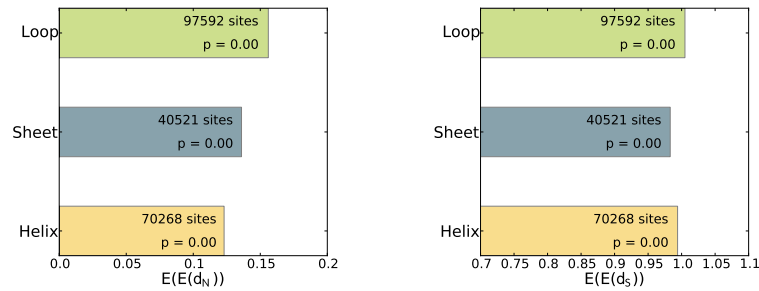
(a) *E. coli* K12, GBDD (+ ψ) dataset



(b) *E. coli* K12, Dual Γ (+ ψ) dataset



(c) *Homo Sapiens*, GBDD (+ ψ) dataset



(d) *Homo sapiens*, Dual Γ (+ ψ) dataset

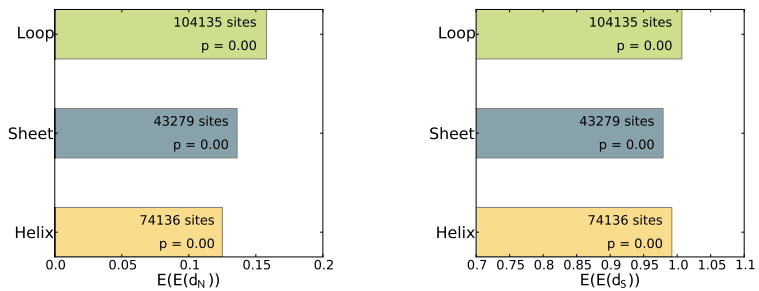


Figure 6.2: The means of the mean posterior rates in the different secondary structures. P -values were obtained by performing a Z -test to measure the significance of the deviation from the global mean rates.

6.2 Correlations between amino acid usage, codon usage and evolutionary rates

On each of the subsets listed in table 6.1 all possible pairwise correlations between the following quantities were investigated for *E. coli K12* and *Homo sapiens*:

- Amino acid frequency
- Absolute codon frequency
- Synonymous codon frequency
- RSCU
- Mean posterior nonsynonymous rate
- Mean posterior synonymous rate

The mean posterior rates investigated here are the mean rates for a site, across the different rate classes, as defined in equation 4.1. Pearson’s correlation coefficient was used to investigate pairwise correlations between the quantities. The correlation coefficient between two sets of values, x and y , is given by:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

If $r = 1$ or $r = -1$ there is a perfect linear relationship between x and y and $r = 0$ indicates no linear relationship. It should be kept in mind that Pearson’s correlation coefficient cannot detect non-linear relationships between variables, hence low values do not rule out the possibility that x and y are related.

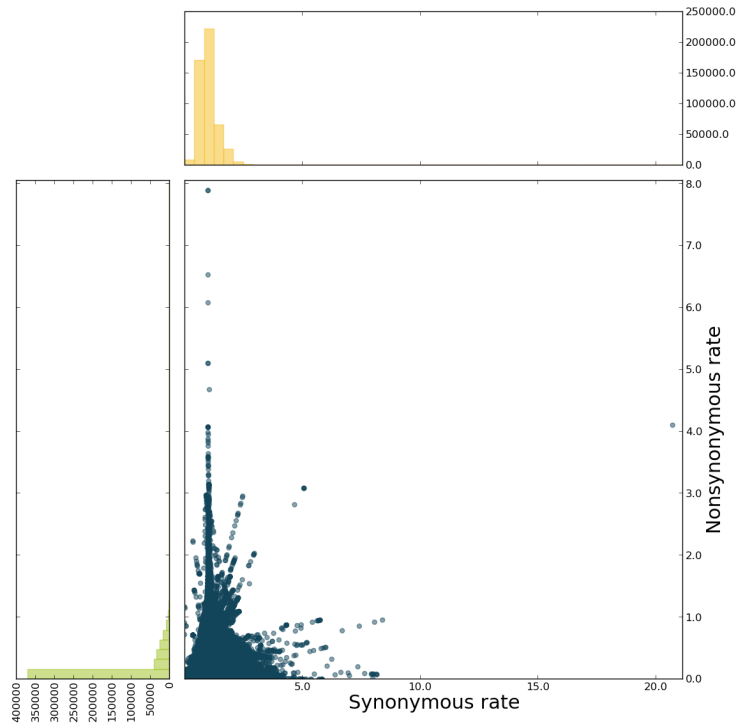
The largest correlations were found between the amino acid frequencies, absolute codon frequencies and synonymous codon frequencies. Large correlations were also found between absolute codon frequencies, synonymous codon frequencies and the RSCU. For all these pairs $|r|$ ranges between 0.3 and 0.6. However, these correlations are expected and should be obvious from the definitions. Interestingly, in most cases $r > 0.1$ for correlations between the amino acid frequency and the RSCU. This shows that the RSCU does not completely remove the amino acid bias in the synonymous codon frequency.

Correlations between the nonsynonymous and synonymous rates and correlations between these rates and frequencies are more interesting. Unfortunately no particularly strong correlations were found between any of these quantities. The only pairs where $|r|$ is greater than 0.1 are between the nonsynonymous and synonymous rates on the GBDD ($r = 0.116$), GBDD + ψ ($r = 0.104$) and GBDD (+ ψ) ($r = 0.11$) datasets and between the synonymous rate and the RSCU on the Dual Γ dataset ($r = -0.118$), all for *Homo sapiens*. The highest correlation between the nonsynonymous and synonymous rates for *E. coli K12* is on the GBDD + ψ dataset ($r = 0.061$). The highest correlation observed between the synonymous rate and the RSCU for *E. coli K12* is on the GBDD model ($r = -0.054$).

Some correlation between the nonsynonymous and synonymous rates is to be expected, since both rates quantify the sequence divergence to some degree. Since sequences are more divergent in the Mammalian dataset than the *E. coli* dataset a higher correlation is expected. It is also not really surprising that the GBDD model shows more correlation between the nonsynonymous and synonymous rates, as the rates are drawn from a bivariate distribution. Looking at the scatter plot for the correlation on *Homo sapiens*, (figure 6.3) a small effect can be observed. However, no effect is observed for the Dual Γ dataset, where the rates are drawn from independent distributions. Hence, it is unlikely that the effect observed on the GBDD dataset is due to any real correlation between the rates and is probably just an artefact of the model.

The correlations between the the synonymous rate and the RSCU are also minimal, and no real effect can be observed in the scatter plots (figure 6.4). It should be noted that although the p -values are highly significant on nearly all pairs, they are inadmissible, as neither the rates nor the frequencies are expected to follow a Gaussian distribution.

(a) *Homo sapiens* GBDD (+ ψ) dataset



(b) *Homo sapiens* Dual Γ (+ ψ) dataset

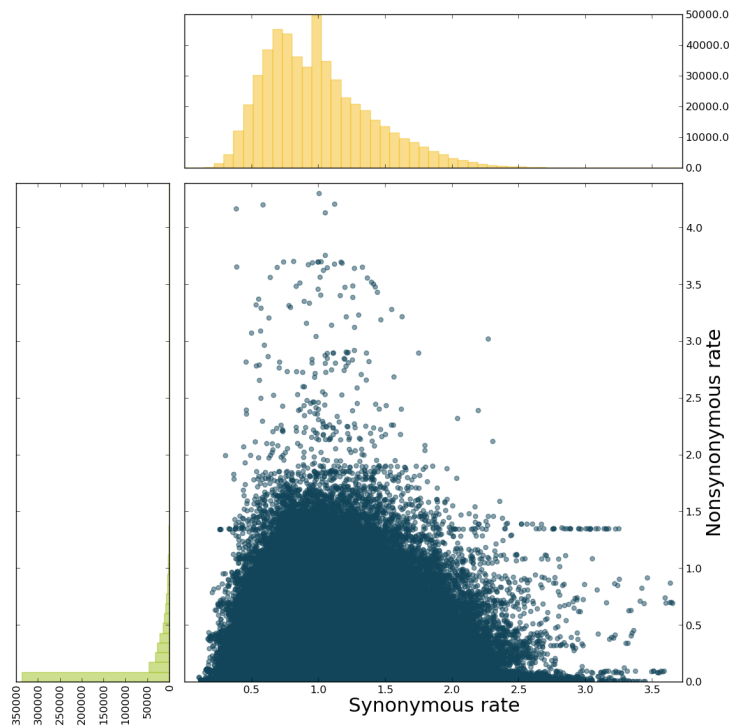
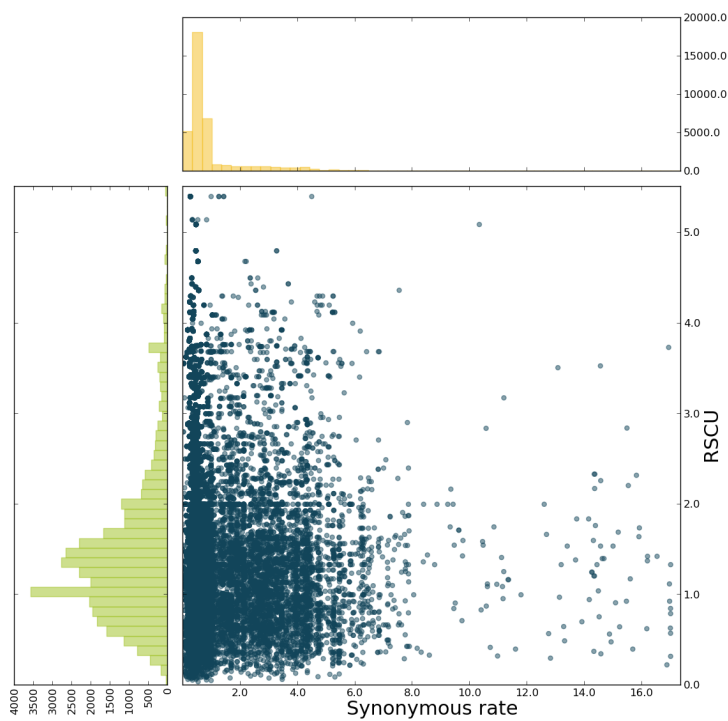


Figure 6.3: Scatter plot of the correlation between the posterior nonsynonymous and synonymous rates over all sites where the respective models were preferred. Note that the axes of the two plots are not to scale.

(a) *E. coli* K12 GBDD dataset



(b) *Homo sapiens* Dual Γ dataset

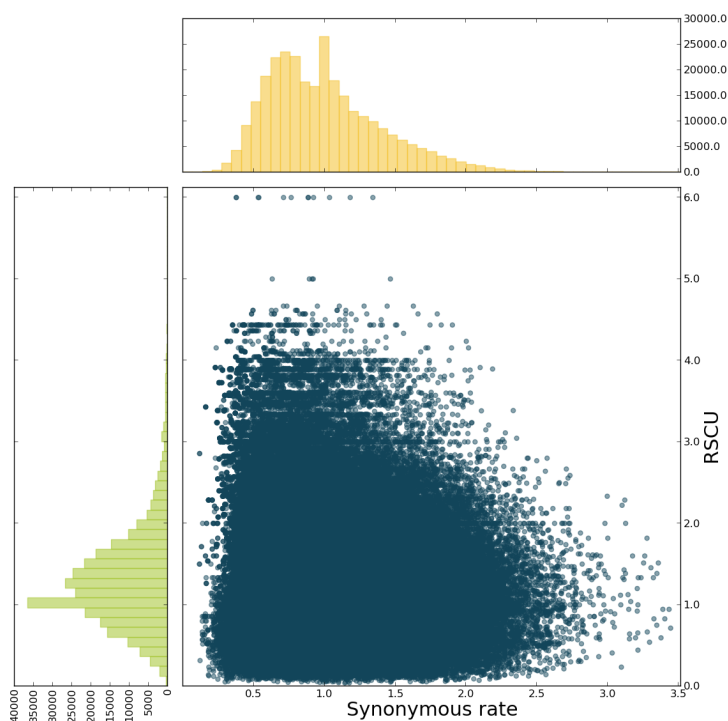


Figure 6.4: Scatter plot of the correlation between the posterior synonymous rate and the RSCU over all sites in all where the respective models were preferred. Note that the axes of the two plots are not to scale.

Table 6.4: The number of sequences that show a significant bias between the posterior evolutionary rates and the codon usage. The datasets are as defined in table 6.1.

	<i>E. coli K12</i>		<i>Homo sapiens</i>	
	$\mathbb{E}(d_N)$	$\mathbb{E}(d_S)$	$\mathbb{E}(d_N)$	$\mathbb{E}(d_S)$
GBDD	77	78	173	197
GBDD + ψ	20	18	174	192
GBDD (+ ψ)	97	96	347	389
Dual Γ	105	103	185	237
Dual Γ + ψ	21	20	148	190
Dual Γ (+ ψ)	126	123	333	427

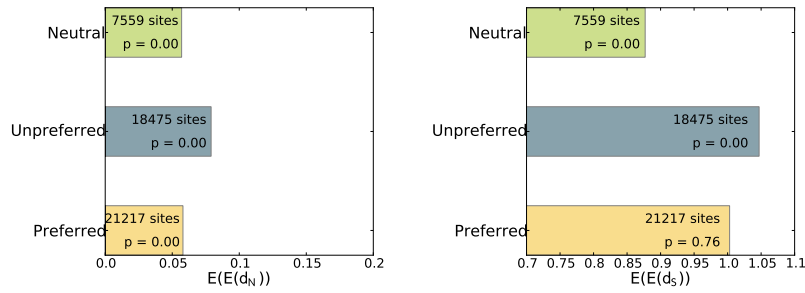
To investigate the relationship between the codon usage and the evolutionary rates more rigorously a χ^2 -test, similar to the ones used in section 6.1 was set up. Using the strategy described in section 5.2.3, each codon was placed into one of three categories: neutral, preferred or unpreferred. Next, the rates were again partitioned into quartiles. A χ^2 -test was then used to assess if rates are preferentially assigned to different codon classes. All tests were performed with $\alpha = 0.05$.

For both organisms all of the datasets tested highly significant. As in section 6.1.2 tests were repeated on individual sequences to find the sequences showing the strongest effects. As can be seen from table 6.4, the effect is very widespread. For *E. coli K12* most of the sequences in all the datasets show an effect. For *Homo sapiens*, the effect is weaker, with less than half of the sequences showing an effect. However, most of the sequences that show an effect for the nonsynonymous rate on *Homo sapiens* also show an effect for the synonymous rate. On the GBDD (+ ψ) dataset the Jaccard index is $341/395$ and on the Dual Γ (+ ψ) dataset it is $330/430$. Although the effect is quite strong, plotting the rate profiles as in figure 6.1 does not show any clear correlations.

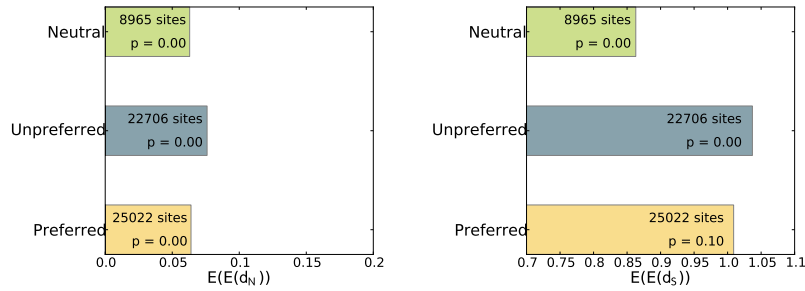
The above results are evidence that preferred and unpreferred codons are under different selective pressures. Furthermore, the results on *Homo sapiens* show that if the synonymous rate varies between preferred and unpreferred codons, then the nonsynonymous rate will also vary with high probability. The variation on the synonymous rate between preferred and unpreferred codons may be ascribed to the selective pressures that act on preferred codons (section 2.2), however it is more challenging to find an explanation for the variation in the nonsynonymous rate. It may be due to a second-order correlation that results in preferred (or unpreferred) codons being preferentially used at sites under purifying (diversifying) selection.

Looking at the means of the mean rates for the different codon classes it can be seen that unpreferred codons always have a higher nonsynonymous and synonymous rate than preferred and neutral codons. (Figure 6.3 shows the results for the GBDD (+ ψ) and Dual Γ (+ ψ) datasets. The results for all the datasets are given in appendix H). It may be postulated that the higher synonymous rate is because unpreferred codons are under higher diversifying selection to mutate. The higher nonsynonymous rate may be because unpreferred codons are preferentially found on diversifying sites, or because they give a fitness advantage and therefore are seen more often. If a site is under positive selection it will not have enough time to optimize its codon usage, hence it is more likely to use unpreferred codons. It is further interesting to note that for *Homo sapiens* the mean synonymous rate on neutral codons is always very close to one, whereas it is significantly lower than the other mean rates for *E. coli K12*.

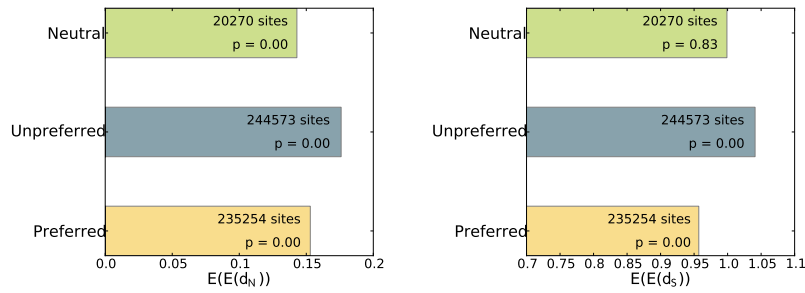
(a) *E. coli K12*, GBDD (+ ψ) dataset



(b) *E. coli K12*, Dual Γ (+ ψ) dataset



(c) *Homo Sapiens*, GBDD (+ ψ) dataset



(d) *Homo sapiens*, Dual Γ (+ ψ) dataset

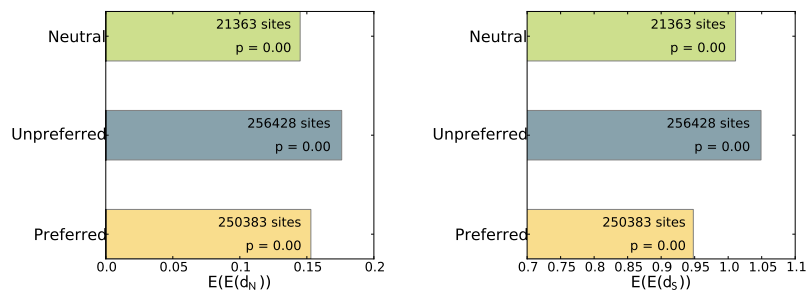


Figure 6.5: The means of the mean posterior rates in the different codon classes. P -values were obtained by performing a Z -test to measure the significance of the deviation from the global mean rates.

6.3 Discussion

In this chapter the relationships between various site-specific quantities were investigated. Although a relationship was detected between secondary structural categories and the usages of a few codons in both organisms, the effect is not a strong one. A much stronger relationship was found between the secondary structural categories and the nonsynonymous and synonymous rates. From the results it can be deduced that different secondary structures are under different evolutionary pressures. Since the effects detected for the nonsynonymous and synonymous rates were found to be mostly independent of each other it may further be deduced that secondary structures are under differential selective constraints both for nonsynonymous and synonymous substitutions. In section 6.1.2, the effect on individual sequences was assessed by repeating the χ^2 -test on each sequence individually. Alternatively, the strategy of Oresic and Shalloway [1998] may be followed. In this strategy the test is performed on the whole dataset, and then on the dataset when neglecting each sequence in turn. In this way the contribution of each sequence to the overall test result can be found, and not overemphasized.

Further, a small correlation was detected between the nonsynonymous and synonymous rates on some datasets. However, this seems to be an artefact of the model that was used to determine the rates, and not a real effect, hence it was not overemphasized. Although the scatter plots do not show a relationship between the RSCU and the evolutionary rates, when the codons are divided into preferred and unpreferred codons a strong effect was found between both evolutionary rates and the codon class. This apparent contradiction may be explained by considering that firstly, codons may be under selection without there being a clear bias [Chamary and Hurst 2005b], and secondly, the assignment of preferred and unpreferred codons was not done on the same dataset, but on a much larger dataset containing all the sequences of the model organism of each dataset in the OMA database (section 5.2.3). In fact, the assignment of preferred and unpreferred codons was repeated using only the sequences in the datasets in question, in which case no preferred codons were detected for either organism. Furthermore, no sequences with a high codon bias were found in either of the datasets. Thus, since none of the sequences show strong codon bias, a correlation between the RSCU and the other quantities is not to be expected. Furthermore, the assignment of codons to preferred and unpreferred classes is independent of the RSCU, hence a low amount of codon bias does not rule out correlations to the codon class.

Chapter 7

Conclusion

The relationships between evolutionary rates, synonymous codon usage and protein secondary structures were investigated for two large datasets of orthologous groups extracted from OMA, a set of *E. coli* strains, and a set of mammalian species. A model selection strategy was followed to identify all groups with significant synonymous variation and selection on the codon usage. An investigation was then carried out to look for effects between the evolutionary rates assigned by models of synonymous variation, the synonymous codon usage and the secondary structural assignments on *E. coli K12* and *Homo sapiens*. Secondary structure assignments were extracted from the PDB.

From the results it can be seen that nonsynonymous and synonymous rate variations are a paired effect. Almost no groups were detected that only showed evidence for variations in the nonsynonymous rate. It was also found that synonymous variation is much more widespread in mammalian genes than in *E. coli* genes. However, the variations are much lower, and probably indicates that only a few sites are under selection. Further, it was found that a significant amount of groups show evidence of both purifying and diversifying selection on the codon usage.

The examination of correlations between the site specific quantities of the sequences of *E. coli* and *Homo sapiens* indicates that for both species there is a non-random distribution of the nonsynonymous and synonymous rates of evolution in different secondary structural classes. On the other hand, only a weak relationship between the synonymous codon usage and the secondary structural classes is present in the data. However, the data show a non-random usage of nonsynonymous and synonymous rates in preferred and unpreferred codons. These results may indicate that relationships between the codon usage, secondary structures and evolutionary rates are more dependent on the bias of a codon within an organism than on the codon bias in the sequence in question. The fact that a strong effect was not detected between the codon usage and the secondary structure does not rule out the possibility that there is a relationship. If there is a relationship it is expected to be weak, and would therefore only be detectable in some cases.

The nature of the significant relationships should be investigated in subsequent research. It is hypothesized that the synonymous rate is lower on preferred codons, as they should be under purifying selection. Similarly, it is hypothesized that the nonsynonymous rate is lower on helices and sheets. These two effects are already visible when looking at the means of the mean posterior rates in different codon classes and structural categories. However, it is expected that the synonymous rate would be lower on loops, where purifying selection may work to preserve clusters of rare codons, but this is not observed in the mean rates. This does not rule out the existence of rare codon clusters, but it does show that if rare codon clusters exist, that they are not widespread.

Appendix A

Validation Results

The results of the validation experiments for the *Drosophila adh*, Flavivirus NS-5 and HIV *vif* datasets are given in this appendix. The results follow the same trends observed in section 4.2.

Table A.1: Validation results for the Constant Rates, Nonsynonymous Γ and Nonsynonymous GDD models.

	Constant Rates			Nonsynonymous Γ			Nonsynonymous GDD		
	$\ell(\hat{\theta}, \mathcal{D}, \mathcal{T})$	κ	ω	$\ell(\hat{\theta}, \mathcal{D}, \mathcal{T})$	κ	$CV(\omega)$	$\ell(\hat{\theta}, \mathcal{D}, \mathcal{T})$	κ	$CV(\omega)$
<i>Drosophila adh</i>									
Goldman-Yang									
CodonPhyML	-4779.7307	1.5821	0.0940	-4612.8438	1.6644	1.2404	-4662.3720	1.6271	1.5583
HyPhy	-4779.7306	1.5822	0.0939	-4670.4021	1.6046	1.2272	-4662.3719	1.6270	1.5582
Muse-Gaut									
CodonPhyML	-4714.2555	1.6610	0.1097	-4611.9664	1.6392	1.2420	-4609.1683	1.6828	1.4590
HyPhy	-4714.2554	1.6611	0.1096	-4612.8437	1.6644	1.2403	-4609.1682	1.6826	1.4592
Yap									
CodonPhyML	-4713.4160	1.6408	0.1088	-4611.9664	1.6392	1.2420	-4608.3816	1.6563	1.4552
Flavivirus NS-5									
Goldman-Yang									
CodonPhyML	-9554.6301	2.2454	0.0250	-9178.0503	2.0933	1.2259	-9187.9087	2.3453	1.7739
HyPhy	-9554.6285	2.2480	0.0249	-9220.3885	2.2815	1.2350	-9187.9079	2.3485	1.7737
Muse-Gaut									
CodonPhyML	-9509.8378	2.0971	0.0339	-9181.1750	2.0494	1.2293	-9147.3587	2.1641	1.7501
HyPhy	-9509.8360	2.1005	0.0337	-9178.0500	2.0941	1.2259	-9147.3583	2.1646	1.7499
Yap									
CodonPhyML	-9513.6037	2.0777	0.0336	-9181.1750	2.0494	1.2293	-9151.8395	2.1159	1.7311
HIV-1 <i>vif</i>									
Goldman-Yang									
CodonPhyML	-3499.5973	3.7234	0.6442	-3371.5545	4.0667	1.1047	-3367.1628	4.1258	1.3960
HyPhy	-3499.5972	3.7253	0.6446	-3377.7014	4.1047	1.0809	-3367.1624	4.1255	1.3965
Muse-Gaut									
CodonPhyML	-3507.3007	3.9390	0.7171	-3375.9149	3.9708	1.1104	-3364.1310	4.0590	1.3853
HyPhy	-3507.3006	3.9398	0.7176	-3371.5544	4.0675	1.1047	-3364.1312	4.0581	1.3853
Yap									
CodonPhyML	-3515.5249	3.8492	0.7136	-3375.9149	3.9708	1.1104	-3368.1182	3.9683	1.4012

Table A.2: Validation results for the Dual Γ and Dual GDD models.

	Dual Γ				Dual GDD				
	$\ell(\hat{\theta}, \mathcal{D}, \mathcal{T})$	κ	$CV(d_N)$	$CV(\omega)$	$\ell(\hat{\theta}, \mathcal{D}, \mathcal{T})$	κ	$CV(d_N)$	$CV(\omega)$	
<i>Drosophila adh</i>									
Goldman-Yang									
CodonPhyML	-4654.8879	1.6439	1.2277	0.5526	1.5307	1.6671	1.5559	0.2308	10.9095
HyPhy	-4654.8878	1.6435	1.2277	0.5524	1.5305	1.6668	1.5559	0.5557	10.9064
Muse-Gaut									
CodonPhyML	-4609.8372	1.6748	1.2407	0.3630	1.3738	1.6932	1.4571	0.1500	1.4830
HyPhy	-4609.8371	1.6747	1.2407	0.3626	1.3736	1.6931	1.4570	0.3446	1.6033
Yap									
CodonPhyML	-4609.1906	1.6509	1.2423	0.3558	1.3704	1.6544	1.3477	0.3343	1.4633
Flavivirus NS-5									
Goldman-Yang									
CodonPhyML	-9182.1012	2.0859	1.2356	0.7101	1.7270	2.2063	1.7416	0.4600	2.4980
HyPhy	-9182.1000	2.0852	1.2356	0.7109	1.7280	2.2062	1.7408	0.4471	2.5903
Muse-Gaut									
CodonPhyML	-9150.7393	2.0290	1.2266	0.6023	1.5845	2.2189	1.7319	0.3392	2.0895
HyPhy	-9150.7384	2.0287	1.2265	0.6032	1.5854	2.2191	1.7316	0.4055	2.3776
Yap									
CodonPhyML	-9200.3533	2.0122	1.4081	0.6261	1.8129	2.1870	1.7154	0.3396	2.0607
HIV-1 <i>vif</i>									
Goldman-Yang									
CodonPhyML	-3358.7199	4.0882	1.0815	0.9044	1.8153	4.0734	1.4022	0.4455	6.4816
HyPhy	-3358.7198	4.0870	1.0814	0.9045	1.8153	4.0671	1.4021	0.9992	1.6447
Muse-Gaut									
CodonPhyML	-3359.1476	3.9676	1.1048	0.8275	1.7358	3.9029	1.3868	0.9668	1.5846
HyPhy	-3359.1474	3.9684	1.1048	0.8277	1.7359	3.9031	1.3865	0.9674	1.5835
Yap									
CodonPhyML	-3363.5811	3.8878	1.1105	0.8250	1.7389	3.8301	1.4006	0.9601	1.5959

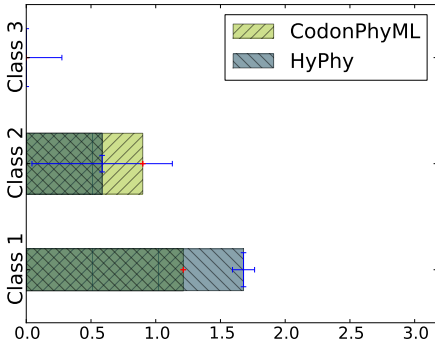
Table A.3: Validation results for the GBDD model.

	$\ell(\hat{\theta}, \mathcal{D}, \mathcal{T})$	κ	$CV(d_N)$	$CV(d_S)$	$CV(\omega)$	
<hr/>						
Goldman-Yang						
CodonPhyML	-4614.8954	1.6242	1.4707	0.1590	1.3658	<i>Drosophila adh</i>
Muse-Gaut						
CodonPhyML	-4592.0149	1.6856	1.4450	0.0324	1.4021	
HyPhy	-4592.0150	1.6857	1.4434	0.1602	1.2529	
Yap						
CodonPhyML	-4590.7235	1.6567	1.4390	0.0393	1.3884	
<hr/>						
Goldman-Yang						
CodonPhyML	-9140.4747	2.2128	1.4492	0.3384	1.4912	Flavivirus NS-5
Muse-Gaut						
CodonPhyML	-9130.1282	2.1959	1.7773	0.0780	1.6042	
HyPhy	-9130.1285	2.1933	1.7626	0.0805	1.6283	
Yap						
CodonPhyML	-9135.0855	2.1288	1.7510	0.0479	1.7058	
<hr/>						
Goldman-Yang						
CodonPhyML	-3345.4462	4.1663	1.3767	0.1632	1.1343	HIV-1 <i>vif</i>
Muse-Gaut						
CodonPhyML	-3341.7262	4.1445	1.3251	0.6734	1.0062	
HyPhy	-3341.7272	4.1461	1.3270	0.6503	0.9175	
Yap						
CodonPhyML	-3348.6388	4.0472	1.3407	0.6834	1.0223	
<hr/>						

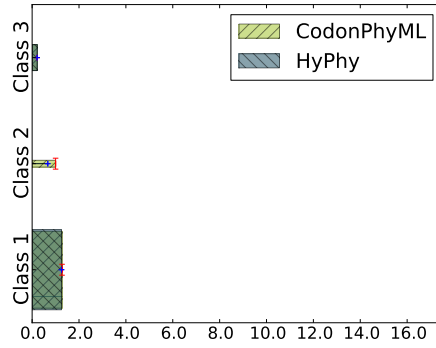
Table A.4: Validation results for the Constant Rates + ψ model.

	$\ell(\hat{\theta}, \mathcal{D}, \mathcal{T})$	κ	ω	ψ		
<hr/>						
Goldman-Yang					<i>Drosophila adh</i>	
CodonPhyML	-4777.0965	1.6234	1.2687	0.1083		
HyPhy	-4777.0966	1.6248	1.2710	0.1084		
<hr/>						
Muse-Gaut						
CodonPhyML	-4708.7174	1.6939	1.3980	0.1329		
<hr/>						
Yap						
CodonPhyML	-4708.4758	1.6705	1.3725	0.1305		
<hr/>						
Goldman-Yang					Flavivirus NS-5	
CodonPhyML	-9554.1834	2.2514	0.9067	0.0237		
HyPhy	-9554.1821	2.2570	0.9080	0.0236		
<hr/>						
Muse-Gaut						
CodonPhyML	-9509.7663	2.0886	1.0419	0.0350		
<hr/>						
Yap						
CodonPhyML	-9513.5612	2.0714	1.0325	0.0344		
<hr/>						
Goldman-Yang					HIV-1 <i>vif</i>	
CodonPhyML	-3499.3359	3.7047	0.8818	0.5915		
HyPhy	-3499.3364	3.6978	0.8779	0.5888		
<hr/>						
Muse-Gaut						
CodonPhyML	-3507.1066	3.9603	1.1132	0.7691		
<hr/>						
Yap						
CodonPhyML	-3515.5002	3.8561	1.0391	0.7319		
<hr/>						

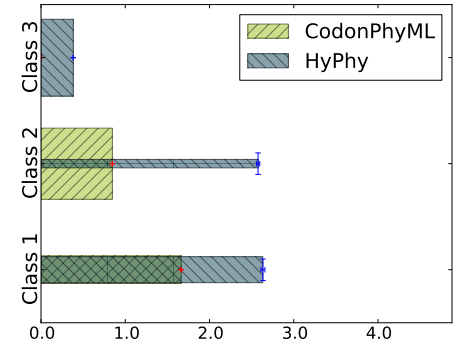
Dual GDD model with Goldman-Yang frequencies



(a) *Drosophila adh*

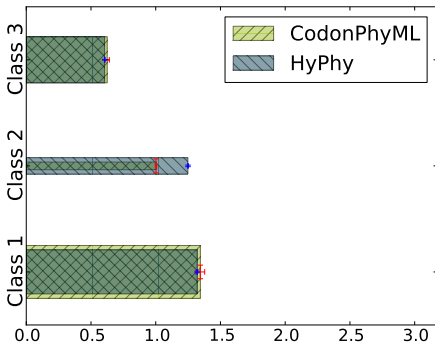


(b) Flavivirus NS-5

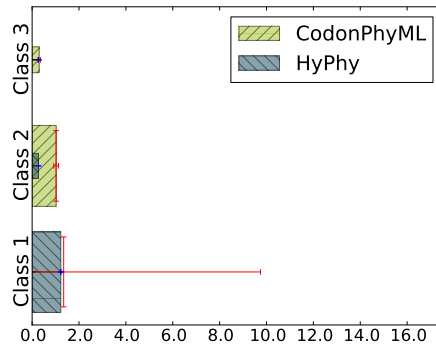


(c) HIV-1 *vif*

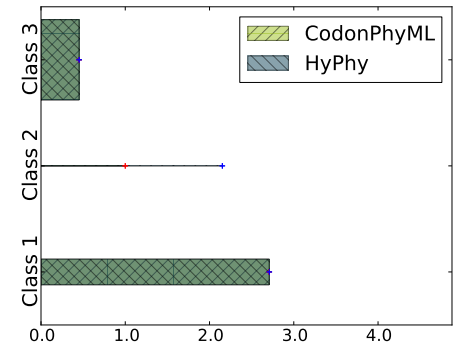
Dual GDD model with Muse-Gaut frequencies



(d) *Drosophila adh*

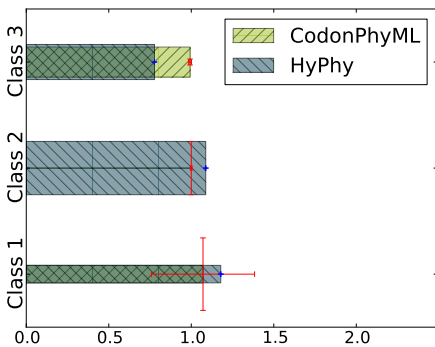


(e) Flavivirus NS-5

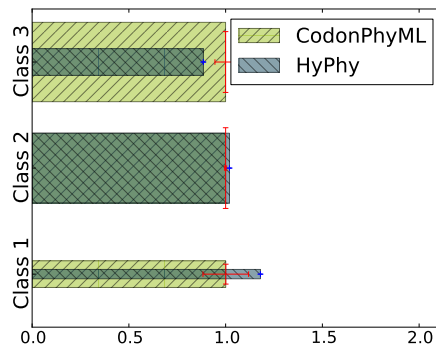


(f) HIV-1 *vif*

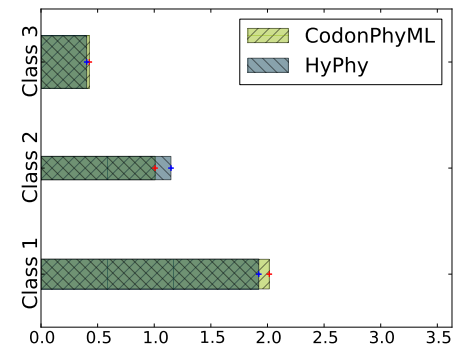
GBDD model with Muse-Gaut frequencies



(g) *Drosophila adh*



(h) Flavivirus NS-5



(i) HIV-1 *vif*

Figure A.1: Comparison of the synonymous rates found using CodonPhyML and HyPhy with the Dual GDD and GBDD models coupled with different frequency models. The width of the bars are proportional to the weight assigned to each rate. The values plotted are for the median rate from 5 runs. The error bars for CodonPhyML and HyPhy are respectively red and blue and were produced from a trimmed sample. (In each case the smallest and largest values were removed).

Appendix B

Extraction of preferred codons

In this appendix the method used to extract preferred codons is examined in more detail. Table B.1 shows the results of varying the selection criteria for *Homo sapiens* and *E. coli K12*. The following three criteria are varied:

- Use the synonymous codon frequency or the absolute codon frequency in the test.
- The level of significance of the χ^2 -test, α .
- The percentage of sequences extracted with the highest and lowest \hat{N}'_c , p .

It can be seen that varying the level of significance does not change which codons are selected. Similarly, varying the percentage of sequences extracted almost never makes a difference. On the other hand, using the absolute codon frequencies results in more codons being selected. Using the synonymous codon frequencies makes more sense intuitively, especially when it is considered that using the absolute codon frequencies makes it possible for stop codons and non-redundant codons (methionine and tryptophan) to be chosen as preferred codons (this is also observed in some cases). Furthermore, absolute codon frequencies do not take the inherent amino acid bias into account.

The assignment strategy for $p = 5\%$ and $\alpha = 0.05$, using synonymous frequencies, is compared to the results of other researchers in table B.2. The methods used by the other researchers are briefly discussed in section 2.1.3. The method described in section 5.2.3 may also be used to assign rare codons. Although it identified far more rare codons than the methods of Wu *et al.* [2010] and Thanaraj and Argos [1996b] it is encouraging that most of the codons identified by these two researchers as rare are also found by the method described here. However, there is a large disagreement between the rare codons identified by the different authors. Further, the method of Thanaraj and Argos [1996b] is not very rigorous, since it only relies on the RSCU of a codon. Hence, the results of Wu *et al.* [2010] are probably more correct. Interestingly, all of the codons found by both authors are also found by the method described here.

From the results in tables B.1 and B.2 it may be concluded that the method is robust on the dataset and the thresholds used. One drawback of the method seems to be that it is not conservative enough in assigning rare codons. The preferred codons found for all the organisms in the *E. coli* and Mammalian datasets, using synonymous frequencies for $p = 5\%$ and $\alpha = 0.05$ are shown in tables B.3 and B.4.

Appendix C

Comparison of frequency models

The model selection experiments reported in tables 5.5, 5.6 and 5.8 were repeated using the Goldman-Yang and Muse-Gaut frequency models. On the *E. coli* dataset optimization failed on 17 groups with the Goldman-Yang model and 20 groups with the Muse-Gaut model. On the Mammalian dataset optimization failed on 10 groups with the Goldman-Yang model and 13 groups with the Muse-Gaut model.

It can be seen from the results that the data follow the same trends as were observed with the Yap frequency model (tables C.1–C.6). However, it can be seen that the Goldman-Yang model prefers more parameter-rich models, whereas the Muse-Gaut model shows results similar to the Yap model. This effect is more pronounced on the Mammalian dataset.

Figure C.1 shows that there is a large overlap between the sets selected with different frequency models coupled to the same model of rate heterogeneity for the Constant Rates, GBDD and Dual Γ models. (Model selection was performed using forward selection). Not enough groups were assigned to the Nonsynonymous GDD and Nonsynonymous Γ models to allow for a large overlap. Furthermore, a fairly large overlap can be seen between the GBDD and Dual Γ models, even between different frequency models.

The distribution of ω values for the Constant Rates model, as well as the nonsynonymous and synonymous coefficients of variation for the GBDD and Dual Γ models were investigated in order to see if there is a consistent bias in one of the frequency models. All three frequency models produced similar histograms with only minor shifts in all of the quantities. These results show that on a dataset like the one that was used here, the frequency model that is used makes no significant difference to the groups that will be selected, or the rates that will be estimated.

Table C.1: Results of the different selection strategies on the *E. coli* dataset for models where rate heterogeneity is treated with a Γ distribution, for different numbers of classes.

Selection Strategy	Constant Rates						Nonsynonymous variation						Nonsynonymous and Synonymous variation					
	Groups Assigned (of 525)	High Support ($\Delta_i \leq 2$)	Jaccard Index wrt AIC	Jaccard Index wrt AIC _c	Jaccard Index wrt BIC	Groups Assigned (of 525)	High Support ($\Delta_i \leq 2$)	Jaccard Index wrt AIC	Jaccard Index wrt AIC _c	Jaccard Index wrt BIC	Groups Assigned (of 525)	High Support ($\Delta_i \leq 2$)	Jaccard Index wrt AIC	Jaccard Index wrt AIC _c	Jaccard Index wrt BIC			
Γ_3 path																		
AIC	105	148				12	46				408	440						
AIC _c	151	185				11	33				363	403						
BIC	217	242				9	18				299	327						
Forward Selection	391		105/391	150/392	212/396	8		7/13	7/12	6/11	126		126/408	126/363	120/305			
Backward Elimination	113		102/116	109/155	113/217	8		7/13	7/12	6/11	404		401/411	363/404	299/404			
Γ_6 path																		
AIC	96	133				15	47				414	447						
AIC _c	134	169				13	33				378	412						
BIC	198	217				12	21				315	345						
Forward Selection	381		96/381	134/381	197/382	9		8/16	9/13	8/13	135		135/414	134/379	131/319			
Backward Elimination	106		95/107	102/138	106/198	9		8/16	9/13	8/13	410		409/415	377/411	315/410			
Γ_9 path																		
AIC	97	130				15	48				413	447						
AIC _c	133	164				16	34				376	412						
BIC	195	212				12	25				318	345						
Forward Selection	354		97/354	133/354	186/363	15		11/19	14/17	8/19	156		156/413	155/377	150/324			
Backward Elimination	101		95/103	100/134	101/195	15		11/19	14/17	8/19	409		407/415	375/410	318/409			

* All tests performed at $\alpha = 0.05$

Table C.2: Results of the different selection strategies on the *E. coli* dataset for models where rate heterogeneity is treated with a Γ distribution, for different numbers of classes.

Selection Strategy	Constant Rates				Nonsynonymous variation				Nonsynonymous and Synonymous variation						
	Groups Assigned (of 522)	High Support ($\Delta_i \leq 2$)	Jaccard Index wrt AIC	Jaccard Index wrt AIC _c	Jaccard Index wrt BIC	Groups Assigned (of 522)	High Support ($\Delta_i \leq 2$)	Jaccard Index wrt AIC	Jaccard Index wrt AIC _c	Jaccard Index wrt BIC	Groups Assigned (of 522)	High Support ($\Delta_i \leq 2$)	Jaccard Index wrt AIC	Jaccard Index wrt AIC _c	Jaccard Index wrt BIC
Γ_3 path															
AIC	125	170				14	56				383	429			
AIC _c	176	198				11	34				335	382			
BIC	241	269				10	26				271	298			
Forward Selection	395		125/395	175/396	235/401	10		7/17	7/14	7/13	117		117/383	117/335	111/277
Backward Elimination	127		116/136	123/180	127/241	10		7/17	7/14	7/13	385		376/392	333/387	271/385
Γ_6 path															
AIC	110	161				18	59				394	433			
AIC _c	165	187				15	41				342	396			
BIC	222	249				14	27				286	312			
Forward Selection	383		110/383	165/383	220/385	11		8/21	9/17	9/16	128		128/394	127/343	123/291
Backward Elimination	120		105/125	116/169	120/222	11		8/21	9/17	9/16	391		386/399	339/394	286/391
Γ_9 path															
AIC	108	152				18	59				396	433			
AIC _c	159	183				19	41				344	397			
BIC	222	245				14	32				286	317			
Forward Selection	360		108/360	159/360	212/370	16		13/21	14/21	9/21	146		146/396	143/347	138/294
Backward Elimination	115		105/118	111/163	115/222	16		13/21	14/21	9/21	391		388/399	340/395	286/391

* All tests performed at $\alpha = 0.05$

Table C.3: Results of the different selection strategies on the *E. coli* dataset for models where rate heterogeneity is treated with a GDD distribution.

Selection Strategy	Constant Rates				Nonsynonymous variation				Nonsynonymous and Synonymous variation					
	Groups Assigned (of 525)	High Support ($\Delta_i \leq 2$)	Jaccard Index wrt AIC	Jaccard Index wrt AIC _c	Jaccard Index wrt AIC	High Support ($\Delta_i \leq 2$)	Jaccard Index wrt AIC	Jaccard Index wrt AIC _c	Jaccard Index wrt BIC	Groups Assigned (of 525)	High Support ($\Delta_i \leq 2$)	Jaccard Index wrt AIC	Jaccard Index wrt AIC _c	Jaccard Index wrt BIC
Dual GDD model														
AIC	222	245				17	26			286	293			
AIC _c	287	297			11	18				227	242			
BIC	400	403			7	10				118	126			
Forward Selection	409		222/409	282/414			11/20	10/15	1/20	102		102/286	100/229	70/150
Backward Elimination	188		181/229	188/287	14		11/20	10/15	1/20	323		282/327	227/323	118/323
GBDD model														
AIC	188	215			13	17				324	349			
AIC _c	243	254			4	9				278	293			
BIC	355	360			2	3				168	181			
Forward Selection	409		188/409	239/413	7		7/13	4/7	1/8	109		109/324	108/279	90/187
Backward Elimination	130		123/195	130/243	7		7/13	4/7	1/8	388		323/389	278/388	168/388

* All tests performed at $\alpha = 0.05$

Table C.4: Results of the different selection strategies on the *E. coli* dataset for models where rate heterogeneity is treated with a GDD distribution.

Selection Strategy	Constant Rates			Nonsynonymous variation			Nonsynonymous and Synonymous variation		
	Groups Assigned (of 522)	High Support ($\Delta_i \leq 2$)	Jaccard Index wrt AIC	Groups Assigned (of 522)	High Support ($\Delta_i \leq 2$)	Jaccard Index wrt AIC	Groups Assigned (of 522)	High Support ($\Delta_i \leq 2$)	Jaccard Index wrt AIC
Dual GDD path									
AIC	251	273		18	33		253	269	
AIC _c	303	316		13	20		206	219	
BIC	414	423		11	14		97	106	
Forward Selection	412		251/412 205/260	12		11/19 11/19	98		98/253 250/299
Backward Elimination	214		214/303	12		11/14 11/14	296		93/211 206/296
GBDD model									
AIC	205	245		12	16		305	329	
AIC _c	262	276		8	10		252	268	
BIC	371	381		3	5		148	158	
Forward Selection	412		205/412 145/209	9		8/13 8/13	101		101/305 304/365
Backward Elimination	149		149/262	9		8/9 8/9	364		100/253 252/364
									60/135 97/296
									84/165 148/364

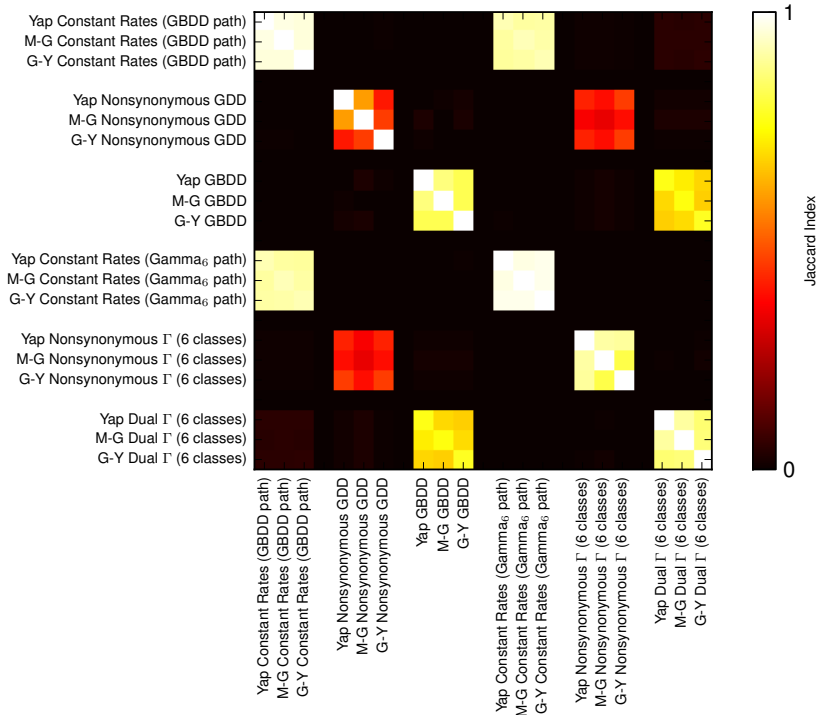
* All tests performed at $\alpha = 0.05$

Table C.5: Results of the different selection strategies on the Mammalian dataset for the two model selection paths that were used to model rate heterogeneity.

Selection Strategy	Constant Rates			Nonsynonymous variation			Nonsynonymous and Synonymous variation			
	Groups Assigned (of 902)	High Support ($\Delta_i \leq 2$)	Jaccard Index wrt AIC	Jaccard Index wrt AIC _c	Jaccard Index wrt BIC	Groups Assigned (of 902)	High Support ($\Delta_i \leq 2$)	Jaccard Index wrt AIC	Jaccard Index wrt AIC _c	Jaccard Index wrt BIC
Γ_6 path										
AIC	0	2				4	6			
AIC _c	2	4			5	9				
BIC	3	4			8	14				
Forward Selection	18	0/18	2/18		6	4/6	5/6	6/8	878/898	876/893
Backward Elimination	0	0/0	0/2	3/18	6	4/6	5/6	6/8	896/898	891/896
				0/3	6					
GBDD model										
AIC	7	7			25	35				
AIC _c	9	10			29	44				
BIC	20	23			59	70				
Forward Selection	32	7/32	8/33	15/37	34	25/34	28/35	31/62	836/870	806/853
Backward Elimination	5	5/7	5/9	5/20	34	25/34	28/35	31/62	861/872	823/863

* All tests performed at $\alpha = 0.05$

(a) *E. coli* dataset



(b) Mammalian dataset

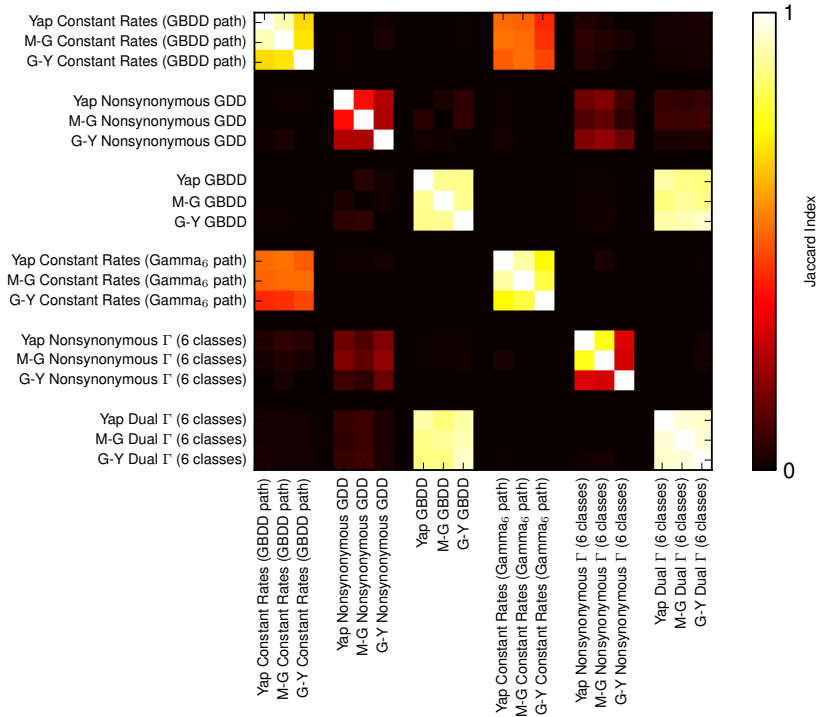


Figure C.1: Heatmap of the Jaccard indices showing the overlap between groups selected using models on the GBDD and Γ_6 paths coupled with different frequency models

Appendix D

OMA groups with evidence of selection on the synonymous codon usage

D.1 *E. coli* dataset

Table D.1: OMA groups in the *E. coli* dataset where positive selection on the codon usage was detected.

OMA group	Group description	ψ
126973	Biotin carboxyl carrier protein	1.26
139417	Ethanolamine utilization protein eutK	2.1
25286	Ribonuclease E	2.14
190	Deaminase	1.54
277981	Ureidoglycolate hydrolase	1.38
43694	Glycerol kinase	1.27
24905	Transcription-repair coupling factor	1.34
40211	D-alanyl-D-alanine carboxypeptidase/D-alanyl-D-alanine-endopeptidase	1.33
123107	4-methyl-5 b-hydroxyethyl -thiazole monophosphate biosynthesis	3.11
36445	Potassium efflux system protein	2.1
263139	Bacterioferritin	1.37
282502	PHP domain protein	1.52
251557	Arginine repressor	1.41
255795	Trna 5-methylaminomethyl-2-thiouridylate - methyl-transferase	1.34

Table D.2: OMA groups in the *E. coli* dataset where negative selection on the codon usage was detected.

OMA group	Group description	ψ
295103	Protein cyaY	0.67
290587	Hydrogenase-2 operon protein hybE	0.65
65152	Bifunctional beta-cystathionase PLP-dependent and regulator of maltose regulon	0.67
246740	Glycogen synthase	0.58
52949	Outer membrane assembly lipoprotein YfgL	0.39
251108	Isocitrate lyase	0.55
1147	33 kDa chaperonin	0.26
256900	FolC bifunctional protein	0.75
107920	Transcriptional regulatory protein ompr	0.62
243019	Aconitate hydratase 2	0.67
92653	Cell division protein ftsQ	0.69
22269	Peptide deformylase	0.39
82908	3-methyl-adenine DNA glycosylase II	0.55
287789	Glutaredoxin 2	0.57
290690	Glutathionylspermidine	0.28
10965	Low molecular weight protein-tyrosine-phosphatase wzb	0.74
1654	Transcriptional activator rfah	0.18
111095	Nitrate/nitrite response regulator protein narL	0.58
285610	Fe-S metabolism associated SufE	0.51
296264	Exonuclease RNase T and DNA polymerase III	0.71
74128	High-affinity zinc uptake system	0.72
3022	KpsF/GutQ family protein	0.61
292543	Putative transcriptional regulator CopG family	0.61
253921	Putative uncharacterized protein	0.66
96121	DNA-binding transcriptional dual regulator	0.56
252770	Diacylglycerol kinase	0.5
284138	Translation initiation factor Sui1	0.75
273922	4-deoxy-l-threo-5-hexosulose-uronate ketol-merase iso-	0.46
277489	Putative exonuclease RdgC	0.3
16810	3-oxoacyl-acyl-carrier-protein synthase	0.67
3091	Phosphoheptose isomerase	0.71
289678	tRNA pseudouridine synthase D	0.6
33500	Transketolase	0.73
246494	Pts system glucose-specific IIA component	0.43
287400	Putative uncharacterized protein	0.15
295701	Methionine synthase	0.56
279726	Extracellular solute-binding protein family 1	0.38
3896	Branched-chain amino acid aminotransferase	0.57
4914	Shikimate kinase	0.4
248799	Methylglyoxal synthase	0.48
295031	Aspartate-ammonia ligase	0.64
62366	Protein tolA	0.55

D.2 Mammalian dataset

Table D.3: OMA groups in the Mammalian dataset where positive selection on the codon usage was detected.

OMA group	Group description	ψ
58221	Protein p48 Chromatin assembly factor 1 subunit C CAF-1 subunit C Chromatin assembly factor I p48 subunit CAF-I 48 kDa subunit CAF-I p48 Nucleoso	1.55
35544	Protein CBFA2T1 Protein MTG8 Protein ETO Eight twenty one protein Cyclin-D-related protein Zinc finger MYND domain- containing protein 2	1.31
420026	VIP peptides precursor Contains Intestinal peptide PHV-42	1.45
45287	T-box transcription factor TBX5 T-box protein 5	1.32
315069	DNA damage-binding protein 2 Damage-specific DNA- binding protein 2	1.4
53244	Ras GTPase-activating protein-binding protein 1	1.26
24652	Son of sevenless homolog 1	2.32
39444	Coactivator-associated arginine methyltransferase 1	1.73
275547	Transaldolase	1.39
41491	Insulin-like growth factor 2 mRNA-binding protein 1 IGF2 mRNA-binding protein 1	1.88
29755	Suppressor of tumorigenicity protein 14	1.3
127883	Chromobox protein homolog 5 Heterochromatin pro- tein 1 homolog alpha	1.31
324614	Fructose-bisphosphate aldolase C	1.33
27407	Protein phosphatase 1 regulatory subunit 12A Myosin phosphatase-targeting subunit 1	1.28
26487	Proto-oncogene tyrosine-protein kinase receptor ret precursor	1.46
47899	Nuclear receptor subfamily 1 group C member 3	1.34
25111	Peregrin Bromodomain and PHD finger-containing protein 1	1.38
359620	Common acute lymphocytic leukemia antigen	1.6
357659	Ubiquitin carrier protein Q1	1.43
368456	Fermitin family homolog 3 Unc-112-related protein 2	1.33
310380	Gamma-butyrobetaine hydroxylase	1.96
316698	Lamin-B receptor Integral nuclear envelope inner membrane protein	1.6
37139	mRNA-capping enzyme HCE HCAP1 Includes Polynucleotide 5'- triphosphatase	1.36
22900	von Willebrand factor precursor vWF Contains von Willebrand antigen 2 von Willebrand antigen II	1.38
297806	Ras GTPase-activating-like protein IQGAP1 p195	1.7

Table D.3: OMA groups in the Mammalian dataset where positive selection on the codon usage was detected (continued).

OMA group	Group description	ψ
313446	CDK-activating kinase assembly factor MAT1 RING finger protein MAT1	1.59
23067	Spectrin beta chain brain 1 Spectrin non- erythroid beta chain 1	1.31
104789	Ankyrin repeat and SOCS box protein 9 ASB-9	1.7
45840	Tyrosine-protein kinase TXK	1.37
91805	Programmed cell death 1 ligand 1 precursor Programmed death ligand 1	1.3
422564	Heat shock protein beta-11 Hspb11	1.62
32870	CD71 antigen Contains Transferrin receptor protein 1 serum form sTfR	1.32
23392	Receptor-type tyrosine-protein phosphatase F Precursor	1.35
73167	Glutaminyl-peptide cyclotransferase precursor	1.46
130684	Ubiquitin-conjugating enzyme E2 F NEDD8 protein ligase UBE2F NEDD8 carrier protein UBE2F	1.41
370857	Regulator of G-protein signaling 19	1.41
121662	Nuclear transcription factor Y subunit beta Nuclear transcription factor Y subunit B	1.45
110191	Adenylate kinase	1.49
73023	Guanine nucleotide-binding protein G k subunit alpha G i alpha-3	1.43
30418	ATP-binding cassette sub-family B member 6 mitochondrial Mitochondrial ABC transporter 3	1.25
33772	Zinc finger and BTB domain-containing protein 27	1.27
32274	Forkhead box protein M1 Forkhead-related protein FKHL16 Hepatocyte nuclear factor 3 forkhead homolog 11 HNF-3/fork-head homolog 11 HFH-11 Winged-helix factor from INS-1 cells M-phase phosphoprotein 2 MPM-2 reactive phosphoprotein 2 Transcri	1.27
306579	Brefeldin A-inhibited guanine nucleotide-exchange protein 2 Brefeldin A-inhibited GEP 2	1.28
115381	Adenylate kinase 3 alpha-like 1	1.37
24239	Nuclear receptor coactivator 1	1.46
44384	Proto-oncogene tyrosine-protein kinase Fyn	1.25
37671	Kinetochores protein NDC80 homolog Kinetochores protein Hec1	1.53
309235	Beta-site APP cleaving enzyme 1 Beta-site amyloid precursor protein cleaving enzyme 1 Membrane-associated aspartic protease 2 Memapsin-2 Aspartyl protease 2 Asp 2 ASP2	1.37
63701	Peptidyl-prolyl cis-trans isomerase PPIase Rotamase 38 kDa FK506-binding protein FKBPR38 hFKBP38	1.39

Table D.3: OMA groups in the Mammalian dataset where positive selection on the codon usage was detected (continued).

OMA group	Group description	ψ
59406	General transcription factor IIE subunit 1 Transcription initiation factor IIE subunit alpha	1.32
97837	Exosome complex exonuclease RRP43	1.43
399094	Transforming growth factor beta regulator 1 Nuclear interactor of ARF and Mdm2	1.3
416443	Growth factor receptor substrate 3 FGFR substrate 3	1.29
29105	Exosome component 10 Polymyositis/scleroderma autoantigen 2 Autoantigen PM/Scl 2 Polymyositis/scleroderma autoantigen 100 kDa PM/Scl-100 P100 polymyositis-scleroderma overlap syndrome-associated autoantigen	1.59
418029	Major vault protein	1.26
125564	Ras-related and estrogen-regulated growth inhibitor	1.28
388108	Transcription factor E2F4 E2F-4	1.29
388109	Transcription factor E2F1 E2F-1	1.46
396242	Transforming growth factor beta-3 precursor	2.93
20776	DNA cross-link repair 1B	1.37
27591	Mast/stem cell growth factor receptor precursor	1.43
438266	Cell death activator CIDE-B Cell death-inducing DFFA-like effector B	1.26
24414	Ribonuclease III	1.38
265174	Acetyl-CoA acetyltransferase mitochondrial precursor	1.76

Table D.4: OMA groups in the Mammalian dataset where negative selection on the codon usage was detected.

OMA group	Group description	ψ
370861	Regulator of G-protein signaling 1	0.65
362264	Fas apoptotic inhibitory molecule	0.68
73164	Protein 63 Muscle-specific RING finger protein 1 MuRF1 MURF-1 RING finger protein 28 Striated muscle RING zinc finger protein Iris RING finger protein	0.74
302037	Proliferating cell nuclear antigen	0.63
41580	Contains Tissue-type plasminogen activator chain A	0.67

Appendix E

OMA groups with a correlation between the nonsynonymous rate and the secondary structure

E.1 *E. coli* K12

Table E.1: OMA groups for *E. coli* sequences from the GBDD (+ ψ) dataset where a correlation between the posterior nonsynonymous rate and the secondary structure was detected.

OMA group	Group description	p -value	χ^2 statistic
75693	Dna polymerase iii	0.00020	26.27182
256361	Oxidoreductase with NAD P - binding Rossmann-fold domain	0.00039	24.70194
267885	Preprotein translocase secA subunit	0.00081	22.95557
290760	Glutamate-ammonia-ligase adenylyltransferase	0.00120	22.03155
36800	Cysteine desulfurase	0.00249	20.26348
9406	Argininosuccinate lyase	0.00451	18.80032
24638	Dehydrogenase	0.00846	17.23352
251910	3-mercaptopyruvate sulfurtransferase	0.00901	17.07644
78	Dna gyrase	0.00927	17.00264
16249	Adenosylmethionine-8-amino-7-oxononanoate aminotransferase	0.01098	16.57420

Table E.2: OMA groups for *E. coli* sequences from the Dual $\Gamma (+\psi)$ dataset where a correlation between the posterior nonsynonymous rate and the secondary structure was detected.

OMA group	Group description	p -value	χ^2 statistic
38167	Glyoxylate carboligase	0.00000	37.80684
286812	D-lactate dehydrogenase	0.00001	32.99925
78	Dna gyrase	0.00001	32.22163
16249	Adenosylmethionine-8-amino-7-oxononanoate aminotransferase	0.00002	32.16622
31966	Glutaminyl-trna synthetase	0.00004	30.11474
50813	Ribosomal RNA small subunit methyltransferase B	0.00004	30.04895
61779	DNA polymerase iii beta	0.00004	29.79671
264954	FeS assembly protein SufD	0.00004	29.71192
34662	Exodeoxyribonuclease v alpha	0.00008	28.41422
39422	Efflux system outer membrane	0.00008	28.39210

E.2 *Homo sapiens*

Table E.3: OMA groups for *Homo sapiens* sequences from the GBDD (+ ψ) dataset where a correlation between the posterior nonsynonymous rate and the secondary structure was detected.

OMA group	Group description	p -value	χ^2 statistic
25106	Apoptotic protease-activating factor 1 Apaf-1	0.00000	58.59921
16936	Tryptophanyl-tRNA synthetase	0.00000	57.06434
131143	Ubiquitin-conjugating enzyme E2 H	0.00000	47.80580
364701	Plexin-B1 precursor Semaphorin receptor SEP	0.00000	42.65576
38827	Coagulation factor II Contains Activation peptide fragment 1	0.00000	42.27467
397050	Talalactoferrin Contains Kaliocin-1	0.00000	39.95954
339876	Neuroserpin precursor Serpin II	0.00000	39.92280
281370	Type 1 tumor necrosis factor receptor shedding aminopeptidase regulator	0.00000	38.39851
259001	DNA- apurinic or apyrimidinic site lyase Neil1	0.00000	35.86479
34962	DNA polymerase iota	0.00000	35.00914

Table E.4: OMA groups for *Homo sapiens* sequences from the Dual $\Gamma (+\psi)$ dataset where a correlation between the posterior nonsynonymous rate and the secondary structure was detected.

OMA group	Group description	p -value	χ^2 statistic
38926	F-box/WD repeat-containing protein 1A F-box and WD repeats protein beta-TrCP	0.00000	60.34451
364701	Plexin-B1 precursor Semaphorin receptor SEP	0.00000	58.30019
76612	DnaJ homolog subfamily B member 1 Heat shock 40 kDa protein 1	0.00000	54.18851
16936	Tryptophanyl-tRNA synthetase	0.00000	48.92961
397050	Talalactoferrin Contains Kaliocin-1	0.00000	45.28118
310591	Neurofibromin Neurofibromatosis-related protein NF-1 Contains Neurofibromin truncated	0.00000	44.43761
25106	Apoptotic protease-activating factor 1 Apaf-1	0.00000	43.20608
281370	Type 1 tumor necrosis factor receptor shedding aminopeptidase regulator	0.00000	41.03581
131143	Ubiquitin-conjugating enzyme E2 H	0.00000	38.77793
38827	Coagulation factor II Contains Activation peptide fragment 1	0.00000	38.55863

Appendix F

OMA groups with a correlation between the synonymous rate and the secondary structure

F.1 *E. coli* K12

Table F.1: OMA groups for *E. coli* sequences from the GBDD (+ ψ) dataset where a correlation between the posterior synonymous rate and the secondary structure was detected.

OMA group	Group description	p -value	χ^2 statistic
34662	Exodeoxyribonuclease v alpha	0.00030	25.32722
286386	Exodeoxyribonuclease I	0.00110	22.22319
290760	Glutamate-ammonia-ligase adenylyltransferase	0.00125	21.92495
75693	Dna polymerase iii	0.00142	21.62587
92787	NAD P H quinone oxidoreductase	0.00754	17.52470
34912	DNA polymerase III	0.00949	16.94527
267885	Preprotein translocase secA subunit	0.01049	16.68964
24905	Transcription-repair coupling factor	0.01252	16.24136
61779	DNA polymerase iii beta	0.01703	15.45025
50169	6-phosphogluconate dehydrogenase decarboxylating	0.02085	14.92441

Table F.2: OMA groups for *E. coli* sequences from the Dual $\Gamma (+\psi)$ dataset where a correlation between the posterior synonymous rate and the secondary structure was detected.

OMA group	Group description	p -value	χ^2 statistic
34662	Exodeoxyribonuclease v alpha	0.00055	23.87927
284020	Tryptophan biosynthesis protein trpcf	0.00062	23.61044
267885	Preprotein translocase secA subunit	0.00078	23.06043
286386	Exodeoxyribonuclease I	0.00141	21.63887
290760	Glutamate-ammonia-ligase adenylyltransferase	0.00203	20.75450
35724	Gamma-glutamyltranspeptidase	0.00278	19.99082
277997	Acyl-CoA synthetase with NAD P-binding Rossmann-fold domain	0.00429	18.92633
255408	O-succinylhomoserine thiol -lyase	0.00520	18.45125
61920	Phosphoribosylglycinamide formyltransferase 2	0.00985	16.84925
83913	Transcriptional regulator LysR family	0.01106	16.55583

F.2 *Homo sapiens*

Table F.3: OMA groups for *Homo sapiens* sequences from the GBDD (+ ψ) dataset where a correlation between the posterior synonymous rate and the secondary structure was detected.

OMA group	Group description	p -value	χ^2 statistic
25106	Apoptotic protease-activating factor 1 Apaf-1	0.00000	45.95978
16936	Tryptophanyl-tRNA synthetase	0.00000	43.37955
259001	DNA- apurinic or apyrimidinic site lyase Neil1	0.00000	38.76860
131143	Ubiquitin-conjugating enzyme E2 H	0.00001	33.79037
357640	Inositol polyphosphate 5-phosphatase OCRL-1	0.00002	31.28552
38926	F-box/WD repeat-containing protein 1A F-box and WD repeats protein beta-TrCP	0.00010	27.78993
49536	Fibrinogen beta chain precursor Contains Fibrinopeptide B	0.00056	23.82116
416709	Plexin-C1 precursor Virus-encoded semaphorin protein receptor	0.00074	23.17699
354621	PAF acetylhydrolase PAF 2-acylhydrolase LDL-associated phospholipase A2 LDL-PLA 2 2-acetyl-1-alkylglycerophosphocholine esterase 1-alkyl-2-acetyl-glycerophosphocholine esterase	0.00133	21.77752
40972	Complement component C8 alpha chain precursor Complement component 8 subunit alpha	0.00159	21.34790

Table F.4: OMA groups for *Homo sapiens* sequences from the Dual $\Gamma (+\psi)$ dataset where a correlation between the posterior synonymous rate and the secondary structure was detected.

OMA group	Group description	p -value	χ^2 statistic
339865	Peptidase inhibitor clade C antithrombin member 1	0.00005	29.57760
357640	Inositol polyphosphate 5-phosphatase OCRL-1	0.00061	23.63579
398682	Apoptotic protease Mch-2 Contains Caspase-6 subunit p18	0.00245	20.29991
27394	Signal transduction protein CBL-B SH3-binding protein CBL-B Casitas B-lineage lymphoma proto-oncogene b RING finger protein 56	0.00249	20.26323
312050	Eukaryotic translation initiation factor 5	0.00314	19.68983
244872	Uridine phosphorylase 1	0.00786	17.41885
48816	Amine oxidase flavin-containing B	0.00867	17.17098
320698	Prostatic acid phosphatase precursor	0.00883	17.12744
287759	Chloride intracellular channel protein 3	0.00913	17.04175
361373	Phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit gamma isoform	0.00914	17.03921

Appendix G

Mean rates in secondary structures

G.1 *E. coli K12*

Table G.1: The means of the mean posterior rates in the different secondary structures, for *E. coli K12*. Significance was assessed by a Z -test. Where the deviation from the total mean rate for the dataset is significant ($p < 0.05$) it is indicated with an *.

		Sites	$\mathbb{E}(\mathbb{E}(d_N))$	p -value	$\mathbb{E}(\mathbb{E}(d_S))$	p -value
GBDD	Helix	12697	0.064	0.024*	1.007	0.569
	Sheet	5887	0.057	0.000*	0.966	0.045*
	Loop	13284	0.077	0.001*	1.009	0.424
GBDD + ψ	Helix	3951	0.044	0.000*	1.008	0.646
	Sheet	1177	0.040	0.006*	0.974	0.402
	Loop	3593	0.050	0.122	0.990	0.579
GBDD (+ ψ)	Helix	16648	0.059	0.000*	1.007	0.478
	Sheet	7064	0.054	0.000*	0.968	0.029*
	Loop	16877	0.071	0.010*	1.005	0.608
Dual Γ	Helix	16086	0.071	0.399	1.003	0.558
	Sheet	7439	0.067	0.009*	0.974	0.088
	Loop	16352	0.076	0.000*	1.018	0.028*
Dual Γ + ψ	Helix	4402	0.051	0.000*	0.995	0.919
	Sheet	1564	0.050	0.000*	0.954	0.106
	Loop	4102	0.053	0.000*	0.989	0.635
Dual Γ (+ ψ)	Helix	20488	0.067	0.039*	1.001	0.626
	Sheet	9003	0.064	0.000*	0.970	0.028*
	Loop	20454	0.071	0.000*	1.012	0.068

G.2 *Homo sapiens*

Table G.2: The means of the mean posterior rates in the different secondary structures for *Homo sapiens*. Significance was assessed by a Z -test. Where the deviation from the total mean rate for the dataset is significant ($p < 0.05$) it is indicated with an *.

		Sites	$\mathbb{E}(\mathbb{E}(d_N))$	p -value	$\mathbb{E}(\mathbb{E}(d_S))$	p -value
GBDD	Helix	37839	0.117	0.000*	0.995	0.023*
	Sheet	23796	0.132	0.000*	0.982	0.000*
	Loop	54623	0.154	0.000*	1.006	0.000*
GBDD + ψ	Helix	32429	0.129	0.000*	0.992	0.000*
	Sheet	16725	0.143	0.000*	0.985	0.000*
	Loop	42969	0.159	0.000*	1.004	0.056
GBDD (+ ψ)	Helix	70268	0.123	0.000*	0.994	0.000*
	Sheet	40521	0.136	0.000*	0.983	0.000*
	Loop	97592	0.156	0.000*	1.005	0.000*
Dual Γ	Helix	44201	0.117	0.000*	0.992	0.000*
	Sheet	27937	0.131	0.000*	0.977	0.000*
	Loop	64905	0.152	0.001*	1.007	0.000*
Dual Γ + ψ	Helix	29935	0.138	0.000*	0.990	0.000*
	Sheet	15342	0.147	0.000*	0.982	0.000*
	Loop	39230	0.166	0.000*	1.005	0.004*
Dual Γ (+ ψ)	Helix	74136	0.125	0.000*	0.992	0.000*
	Sheet	43279	0.136	0.000*	0.979	0.000*
	Loop	104135	0.158	0.000*	1.007	0.000*

Appendix H

Mean rates in codon classes

H.1 *E. coli* K12

Table H.1: The means of the mean posterior rates in the different codon classes, for *E. coli* K12. Significance was assessed by a Z -test. Where the deviation from the total mean rate for the dataset is significant ($p < 0.05$) it is indicated with an *.

		Sites	$\mathbb{E}(\mathbb{E}(d_N))$	p -value	$\mathbb{E}(\mathbb{E}(d_S))$	p -value
GBDD	Preferred	16354	0.060	0.000*	0.998	0.808
	Unpreferred	14441	0.083	0.000*	1.053	0.000*
	Neutral	5946	0.059	0.004*	0.878	0.000*
GBDD + ψ	Preferred	4863	0.049	0.051	1.020	0.206
	Unpreferred	4034	0.063	0.005*	1.026	0.125
	Neutral	1613	0.050	0.283	0.876	0.000*
GBDD (+ ψ)	Preferred	21217	0.058	0.000*	1.003	0.763
	Unpreferred	18475	0.079	0.000*	1.047	0.000*
	Neutral	7559	0.057	0.002*	0.877	0.000*
Dual Γ	Preferred	19483	0.066	0.000*	1.010	0.122
	Unpreferred	18170	0.078	0.000*	1.033	0.000*
	Neutral	7016	0.065	0.000*	0.867	0.000*
Dual Γ + ψ	Preferred	5539	0.057	0.008*	1.004	0.597
	Unpreferred	4536	0.068	0.000*	1.051	0.001*
	Neutral	1949	0.054	0.003*	0.849	0.000*
Dual Γ (+ ψ)	Preferred	25022	0.064	0.000*	1.009	0.104
	Unpreferred	22706	0.076	0.000*	1.037	0.000*
	Neutral	8965	0.063	0.000*	0.863	0.000*

H.2 *Homo sapiens*

Table H.2: The means of the mean posterior rates in the different codon classes, for *Homo sapiens*. Significance was assessed by a Z -test. Where the deviation from the total mean rate for the dataset is significant ($p < 0.05$) it is indicated with an *.

		Sites	$\mathbb{E}(\mathbb{E}(d_N))$	p -value	$\mathbb{E}(\mathbb{E}(d_S))$	p -value
GBDD	Preferred	127332	0.145	0.000*	0.966	0.000*
	Unpreferred	113924	0.158	0.000*	1.038	0.000*
	Neutral	10237	0.136	0.000*	0.999	0.800
GBDD + ψ	Preferred	107922	0.161	0.000*	0.947	0.000*
	Unpreferred	130649	0.191	0.000*	1.044	0.000*
	Neutral	10033	0.151	0.000*	1.000	0.960
GBDD (+ ψ)	Preferred	235254	0.153	0.000*	0.957	0.000*
	Unpreferred	244573	0.176	0.000*	1.041	0.000*
	Neutral	20270	0.143	0.000*	0.999	0.831
Dual Γ	Preferred	152055	0.143	0.000*	0.958	0.000*
	Unpreferred	128507	0.158	0.000*	1.047	0.000*
	Neutral	11873	0.134	0.000*	1.010	0.005*
Dual Γ + ψ	Preferred	98328	0.170	0.000*	0.932	0.000*
	Unpreferred	127921	0.195	0.000*	1.050	0.000*
	Neutral	9490	0.158	0.000*	1.012	0.002*
Dual Γ (+ ψ)	Preferred	250383	0.153	0.000*	0.948	0.000*
	Unpreferred	256428	0.176	0.000*	1.049	0.000*
	Neutral	21363	0.145	0.000*	1.011	0.000*

Bibliography

- Alexei A. Adzhubei, Ivan A. Adzhubeib, Igor A. Krashennnikov, and Stephen Neidle. Non-random usage of degenerate codons is related to protein three-dimensional structure. *FEBS Letters*, 399(1-2):78 – 82, 1996.
- H. Akashi. Synonymous Codon Usage in *Drosophila melanogaster*: Natural Selection and Translational Accuracy. *Genetics*, 136(3):927–935, 1994.
- H. Akashi. Inferring Weak Selection From Patterns of Polymorphism and Divergence at “Silent” Sites in *Drosophila* DNA. *Genetics*, 139(2):1067–1076, 1995.
- Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell*. Garland Science, fifth edition, 2007.
- Adrian M. Altenhoff, Adrian Schneider, Gaston H. Gonnet, and Christophe Dessimoz. OMA 2011: orthology inference among 1000 complete genomes. *Nucleic Acids Research*, 39(suppl 1):D289–D294, 2011.
- Christian B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(4096):223–230, 1973.
- Maria Anisimova and Carolin Kosiol. Investigating protein-coding sequence evolution with probabilistic codon substitution models. *Molecular Biology and Evolution*, 26(2):255–271, 2009.
- Maria Anisimova, Joseph P. Bielawski, and Ziheng Yang. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Molecular Biology and Evolution*, 18(8):1585–1592, 2001.
- Maria Anisimova, Manuel Gil, Jean-Francois Dufayard, Christophe Dessimoz, and Olivier Gascuel. Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. *Systematic Biology*, 2011.
- Le Bao, Hong Gu, Katherine Dunn, and Joseph Bielawski. Methods for selecting fixed-effect models for heterogeneous codon evolution, with comments on their application to gene and genome data. *BMC Evolutionary Biology*, 7:1–13, 2007.
- Soren Brunak and Jacob Engelbrecht. Protein structure and the sequential structure of mRNA: alpha-Helix and beta-sheet signals at the nucleotide level. *Proteins: Structure, Function, and Bioinformatics*, 25(2):237–252, 1996.
- Kenneth P. Burnham and David R. Anderson. *Model Selection and Multimodel Inference: A practical information-theoretic approach*. Springer, second edition, 2002.
- Jean-Vincent Chamary and Laurence D. Hurst. Biased codon usage near intron-exon junctions: selection on splicing enhancers, splice-site recognition or something else? *Trends in Genetics*, 21(5):256 – 259, 2005.
- JV Chamary and Laurence D Hurst. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. 2005.

- J. V. Chamary, Joanna L. Parmley, and Laurence D. Hurst. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat Rev Genet*, 7:98–108, February 2006.
- Maria Luisa Chiusano, Fernando Alvarez-Valin, Massimo Di Giulio, Giuseppe D’Onofrio, Gaetano Ammirato, Giovanni Colonna, and Giorgio Bernardi. Second codon positions of genes and the secondary structures of proteins. relationships and implications for the origin of the genetic code. *Gene*, 261(1):63 – 69, 2000.
- Peter Y. Chou and Gerald D. Fasman. Prediction of protein conformation. *Biochemistry*, 13(2):222–245, 1974. PMID: 4358940.
- Peter J. A. Cock, Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J. L. de Hoon. Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009.
- Josep M. Comeron and Montserrat Aguad. An evaluation of measures of synonymous codon usage bias. *Journal of Molecular Evolution*, 47:268–274, 1998.
- Patricia Cortazzo, Carlos Cerveansky, Mnica Marn, Claude Reiss, Ricardo Ehrlich, and Atilio Deana. Silent mutations affect in vivo protein folding in *Escherichia coli*. *Biochemical and Biophysical Research Communications*, 293(1):537 – 541, 2002.
- Tanya Crombie, Jonathan C. Swaffield, and Alistair J. P. Brown. Protein folding within the cell is influenced by controlled rates of polypeptide elongation. *Journal of Molecular Biology*, 228(1):7 – 12, 1992.
- Valerie Daggett and Alan Fersht. The present view of the mechanism of protein folding. *Nature Reviews Molecular Cell Biology*, 4:497–502, 2003.
- Wayne Delport, Konrad Scheffler, Gordon Botha, Mike B. Gravenor, Spencer V. Muse, and Sergei Kosakovsky Pond. CodonTest: Modeling Amino Acid Substitution Preferences in Coding Sequences. *PLoS Comput Biol*, 6(8):e1000885, 08 2010.
- Christophe Dessimoz and Manuel Gil. Phylogenetic assessment of alignments reveals neglected tree signal in gaps. *Genome Biology*, 11(4):R37, 2010.
- M. Dimitrieva and M. Anisimova. *in preparation*, 2011.
- Joseph Felsenstein. Evolutionary trees from dna sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–376, 1981. 10.1007/BF01734359.
- B S Gaut, B R Morton, B C McCaig, and M T Clegg. Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene *Adh* parallel rate differences at the plastid gene *rbcL*. *Proceedings of the National Academy of Sciences of the United States of America*, 93(19):10274–10279, 1996.
- Nick Goldman and Ziheng Yang. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution*, 11(5):725–736, 1994.
- Wanjun Gu, Tong Zhou, Jianmin Ma, Xiao Sun, and Zuhong Lu. The relationship between synonymous codon usage and protein structure in *Escherichia coli* and *Homo sapiens*. *Biosystems*, 73(2):89 – 97, 2004.
- Stéphane Guindon and Olivier Gascuel. A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Systematic Biology*, 52(5):696–704, 2003.
- Stéphane Guindon, Jean-Francois Dufayard, Vincent Lefort, Maria Anisimova, Wim Hordijk, and Olivier Gascuel. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology*, 59(3):307–321, 2010.

- Yves Guisez, Johan Robbens, Erik Remaut, and Walter Fiers. Folding of the MS2 Coat Protein in *Escherichia coli* is Modulated by Translational Pauses Resulting from mRNA Secondary Structure and Codon Usage: A Hypothesis. *Journal of Theoretical Biology*, 162(2):243 – 252, 1993.
- S. K. Gupta, S. Majumdar, T. K. Bhattacharya, and T. C. Ghosh. Studies on the relationships between the synonymous codon usage and protein secondary structural units. *Biochemical and Biophysical Research Communications*, 269(3):692–696, 2000.
- Ruth Hershberg and Dmitri A. Petrov. Selection on codon bias. *Annual Review of Genetics*, 42(1):287–299, 2008.
- Par Ingvarsson. Molecular evolution of synonymous codon usage in populus. *BMC Evolutionary Biology*, 8(1):307, 2008.
- Kazutaka Katoh and Hiroyuki Toh. Recent developments in the MAFFT multiple sequence alignment program. *Briefings in Bioinformatics*, 9(4):286–298, 2008.
- Rob Knight, Peter Maxwell, Amanda Birmingham, Jason Carnes, J Gregory Caporaso, Brett Easton, Michael Eaton, Micah Hamady, Helen Lindsay, Zongzhi Liu, Catherine Lozupone, Daniel McDonald, Michael Robeson, Raymond Sammut, Sandra Smit, Matthew Wakefield, Jeremy Widmann, Shandy Wikman, Stephanie Wilson, Hua Ying, and Gavin Huttley. PyCogent: a toolkit for making sense from sequence. *Genome Biology*, 8(8):R171, 2007.
- Anton A. Komar and Rainer Jaenicke. Kinetics of translation of [gamma]b crystallin and its circularly permuted variant in an in vitro cell-free system: possible relations to codon distribution and protein folding. *FEBS Letters*, 376(3):195 – 198, 1995.
- Anton A. Komar, Thierry Lesnik, and Claude Reiss. Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. *FEBS Letters*, 462(3):387 – 391, 1999.
- Anton A. Komar. A pause for thought along the co-translational folding pathway. *Trends in Biochemical Sciences*, 34(1):16–24, 2009.
- Sergei Kosakovsky Pond and Simon D. W. Frost. A Simple Hierarchical Approach to Modeling Distributions of Substitution Rates. *Molecular Biology and Evolution*, 22(2):223–234, 2005.
- Sergei Kosakovsky Pond and Simon D. W. Frost. Not so different after all: A comparison of methods for detecting amino acid sites under selection. *Molecular Biology and Evolution*, 22(5):1208–1222, 2005.
- Sergei Kosakovsky Pond and Spencer V. Muse. Site-to-Site Variation of Synonymous Substitution Rates. *Molecular Biology and Evolution*, 22(12):2375–2385, 2005.
- Sergei Kosakovsky Pond, Simon D.W. Frost, and Spencer V. Muse. HyPhy: Hypothesis testing using phylogenies. *Bioinformatics*, 21(5):676–679, 2005.
- Sergei Kosakovsky Pond, Wayne Delport, Spencer V. Muse, and Konrad Scheffler. Correcting the bias of empirical frequency parameter estimators in codon models. *PLoS ONE*, 5(7):e11230, 07 2010.
- Sergei Kosakovsky Pond, Konrad Scheffler, Michael B. Gravenor, Art F.Y. Poon, and Simon D.W. Frost. Evolutionary Fingerprinting of Genes. *Molecular Biology and Evolution*, 27(3):520–536, 2010.
- Carolin Kosiol, Ian Holmes, and Nick Goldman. An empirical codon model for protein sequence evolution. *Molecular Biology and Evolution*, 24(7):1464–1479, 2007.
- Igor A. Krasheninnikov, Anton A. Komar, and Ivan A. Adzhubei. Nonuniform size distribution of nascent globin peptides, evidence for pause localization sites, and a cotranslational protein-folding model. *Journal of Protein Chemistry*, 10:445–453, 1991.
- Ivica Letunic and Peer Bork. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, 23(1):127–128, 2007.

- Ivica Letunic and Peer Bork. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Research*, 2011.
- Helen Lindsay, Von Yap, Hua Ying, and Gavin Huttley. Pitfalls of the most commonly used models of context dependent substitution. *Biology Direct*, 3(1):52, 2008.
- Oksana Lukjancenko, Trudy Wassenaar, and David Ussery. Comparison of 61 Sequenced Escherichia coli Genomes. *Microbial Ecology*, 60:708–720, 2010.
- Cameel H. Makhoul and Edward N. Trifonov. Distribution of rare triplets along mRNA and their relation to protein folding. *Journal of Biomolecular Structure & Dynamics*, 20(3):413–420, 2002.
- Pamela Mukhopadhyay, Surajit Basak, and Tapash Ghosh. Synonymous codon usage in different protein secondary structural classes of human genes: Implication for increased non-randomness of GC3 rich genes towards protein stability. *Journal of Biosciences*, 32:947–963, 2007.
- S V Muse and B S Gaut. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution*, 11(5):715–724, 1994.
- Yasukazu Nakamura, Takashi Gojobori, and Toshimichi Ikemura. Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Research*, 28(1):292, 2000.
- Rasmus Nielsen and Ziheng Yang. Likelihood Models for Detecting Positively Selected Amino Acid Sites and Applications to the HIV-1 Envelope Gene. *Genetics*, 148(3):929–936, 1998.
- Rasmus Nielsen and Ziheng Yang. Estimating the Distribution of Selection Coefficients from Phylogenetic Data with Applications to Mitochondrial and Viral DNA. *Molecular Biology and Evolution*, 20(8):1231–1239, 2003.
- Rasmus Nielsen, Vanessa L. Bauer DuMont, Melissa J. Hubisz, and Charles F. Aquadro. Maximum Likelihood Estimation of Ancestral Codon Usage Bias Parameters in Drosophila. *Molecular Biology and Evolution*, 24(1):228–235, 2007.
- John A. Novembre. Accounting for background nucleotide composition when measuring codon usage bias. *Molecular Biology and Evolution*, 19(8):1390–1394, 2002.
- Matej Oresic and David Shalloway. Specific correlations between relative synonymous codon usage and protein secondary structure. *Journal of Molecular Biology*, 281(1):31 – 48, 1998.
- Joanna L. Parmley and Laurence D. Hurst. Exonic splicing regulatory elements skew synonymous codon usage near intron-exon boundaries in mammals. *Molecular Biology and Evolution*, 24(8):1600–1603, 2007.
- Joanna L. Parmley, J. V. Chamary, and Laurence D. Hurst. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Molecular Biology and Evolution*, 23(2):301–309, 2006.
- John F. Peden. *Analysis of Codon Usage*. PhD thesis, University of Nottingham, 1999.
- Joshua B. Plotkin and Grzegorz Kudla. Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev Genet*, 12(1):32–42, 2011.
- David Posada and Thomas R. Buckley. Model selection and model averaging in phylogenetics: Advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Systematic Biology*, 53(5):793–808, 2004.
- Ian J. Purvis, Andrew J. E. Bettany, T. Chinnappan Santiago, John R. Coggins, Kenneth Duncan, Robert Eason, and Alistair J. P. Brown. The efficiency of folding of some proteins is increased by controlled rates of translation in vivo: A hypothesis. *Journal of Molecular Biology*, 193(2):413 – 417, 1987.

- Nicolas Rodrigue, Nicolas Lartillot, and Herve Philippe. Bayesian comparisons of codon substitution models. *Genetics*, 180(3):1579–1591, 2008.
- Fredrik Ronquist and John P. Huelsenbeck. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12):1572–1574, 2003.
- Adrian Schneider, Gina Cannarozzi, and Gaston Gonnet. Empirical codon substitution matrix. *BMC Bioinformatics*, 6(1):134, 2005.
- Adrian Schneider, Christophe Dessimoz, and Gaston H. Gonnet. OMA Browser: Exploring orthologous relations across 352 complete genomes. *Bioinformatics*, 23(16):2180–2182, 2007.
- D C Shields, P M Sharp, D G Higgins, and F Wright. Silent sites in Drosophila genes are not neutral: evidence of selection among synonymous codons. *Molecular Biology and Evolution*, 5(6):704–716, 1988.
- Donald B. Smith and Peter Simmonds. Characteristics of nucleotide substitution in the Hepatitis C virus genome: Constraints on sequence change in coding regions at both ends of the genome. *Journal of Molecular Evolution*, 45:238–246, 1997.
- Y Suzuki and T Gojobori. A method for detecting positive selection at single amino acid sites. *Molecular Biology and Evolution*, 16(10):1315–1328, 1999.
- Xie Tao and Ding Dafu. The relationship between synonymous codon usage and protein structure. *FEBS Letters*, 434(1-2):93 – 96, 1998.
- T. A. Thanaraj and Patrick Argos. Protein secondary structural types are differentially coded on messenger RNA. *Protein Science*, 5:1973–1983, 1996.
- T. A. Thanaraj and Patrick Argos. Ribosome-mediated translational pause and protein domain organization. *Protein Science*, 5:1594–1612, 1996.
- Julie D. Thompson, Benjamin Linard, Odile Lecompte, and Olivier Poch. A comprehensive benchmark study of multiple sequence alignment methods: Current challenges and future perspectives. *PLoS ONE*, 6(3):e18093, 03 2011.
- Araxi O. Urrutia and Laurence D. Hurst. Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. *Genetics*, 159(3):1191–1199, 2001.
- Frank Wright. The ‘effective number of codons’ used in a gene. *Gene*, 87(1):23 – 29, 1990.
- Xianming Wu, Songfeng Wu, Dong Li, Jiyang Zhang, Lin Hou, Jie Ma, Wanlin Liu, Daming Ren, Yumping Zhu, and Fuchu He. Computational identification of rare codons of Escherichia coli based on codon pairs preference. *BMC Bioinformatics*, 11(1):61, 2010.
- Ziheng Yang and Rasmus Nielsen. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *Journal of Molecular Evolution*, 46:409–418, 1998. 10.1007/PL00006320.
- Ziheng Yang and Rasmus Nielsen. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Molecular Biology and Evolution*, 25(3):568–579, 2008.
- Ziheng Yang, Rasmus Nielsen, Nick Goldman, and Anne-Mette Krabbe Pedersen. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, 155(1):431–449, 2000.
- Ziheng Yang, Wendy S.W. Wong, and Rasmus Nielsen. Bayes empirical bayes inference of amino acid sites under positive selection. *Molecular Biology and Evolution*, 22(4):1107–1118, 2005.
- Ziheng Yang. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution*, 10(6):1396–1401, 1993.
- Ziheng Yang. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *Journal of Molecular Evolution*, 39:306–314, 1994.

- Ziheng Yang. PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer applications in the biosciences : CABIOS*, 13(5):555–556, 1997.
- Ziheng Yang. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution*, 24(8):1586–1591, 2007.
- Von Bing Yap, Helen Lindsay, Simon Easteal, and Gavin Huttley. Estimates of the Effect of Natural Selection on Protein-Coding Content. *Molecular Biology and Evolution*, 27(3):726–734, 2010.
- Marcelo Zanetti, Manuel Gil, and Maria Anisimova. CodonPHYML: Maximum likelihood phylogenetic inference under codon models. *in preparation*, 2011.
- Marcelo Zanetti. *Markovian Codon Models of Evolution for Phylogeny Inference*. Master’s thesis, ETH Zürich, 2010.
- Tong Zhou, Wanjun Gu, and Claus O. Wilke. Detecting positive and purifying selection at synonymous sites in yeast and worm. *Molecular Biology and Evolution*, 2010.