The effect of evolutionary rate estimation methods on correlations observed between substitution rates in models of evolution

by
Stephen Gordon Botha

Submitted in fulfillment towards the degree MSc Computer Science in Mathematical Sciences at the University of Stellenbosch

Supervisor:
Dr K. Scheffler

Date:
17 October 2011

Department of Computer Science

Stellenbosch University

Declaration regarding Plagiarism

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the owner of the copyright thereof (unless to the extent explicitly otherwise stated) and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Signature: ..... S.G. Botha

Date: 17/10/2011

Abstract

The use of the ratio $\omega = \frac{dN}{dS}$ as an indicator of positive selection when modeling evolutionary events within genomes has become widespread, and is used under the assumption that synonymous substitutions are neutral with regards to the evolutionary fitness of a gene, and hence that dS is the substitution rate in the absence of selection. Wyckoff et al (2005) found a strong positive correlation between dS and ω in mammalian genes, which implied that dS was positively correlated with selection acting on a gene – a direct violation of this assumption. Drummond et al. (2008) proposed a hypothesis of selection against mistranslation induced protein misfolding which was shown to remove the effect.

Our study investigates whether the positive correlation observed between dS and ω could be an artifact of the methods used to estimate the values of dN and dS. We also explore whether different model parametrisations have an effect on the correlation. Ascertaining whether the correlation between dS and ω is due to a biological trait, which could indicate that further research is needed into the relationship between evolutionary rates, or whether the correlation is due to a statistical artifact, in which case researchers need to be mindful of the implications of certain methods, could be an important finding for researchers aiming to identify genome regions under positive selection.

We fitted a constant ω model and a site-to-site ω variation model to the data set of Drummond et al. (2008). We investigated the effect of different codon frequency estimation methods, as well as different methods of calculating dN and dS, on the correlation between dS and ω. We found that the positive correlation observed between dS and ω is affected by model parametrisation. In particular, changing the way in which synonymous and nonsynonymous sites are defined and the method of calculating equilibrium codon frequencies not only reduced the positive correlation between dS and ω, but in many cases this correlation became negative. Allowing for ω variation between sites within a gene also had an effect on the correlation between dS and ω in the empirical data.

Our results extend previous work which showed that different models of evolution affect the positive correlation between dS and ω. We have shown which specific model parametrisations and estimation methods could be the cause of disparate correlation patterns between models. Our study indicates that the positive correlation between dS and ω could possibly be artifactual in nature and should probably not be interpreted as biologically relevant until the relationship between dN and dS is better understood.

Uittreksel

Die toepassing van die verhouding $\omega=\frac{dN}{dS}$ as 'n aanwyser van positiewe seleksie wanneer evolusionêre gebeurtenisse in genome gemodelleer word, word algemeen aanvaar en word gebruik met die veronderstelling dat sinverwante plaasvervangers neutraal is met betrekking tot die evolusionêre geskiktheid van 'n geen. Derhalwe is dS die plaasvervangingstempo in die afwesigheid van seleksie. Wyckoff et al (2005) het 'n sterk positiewe korrelasie gevind tussen dS en ω in soogdiergene, wat impliseer dat dS positief gekorreleer is met seleksie wat inwerk op 'n geen – direk teenstrydig met bogenoemde veronderstelling. Drummond et al (2008) het 'n hipotese van seleksie teen proteïenwanvouding geinduseer deur wanvertaaling voorgestel. Hierdie hipotese het geblyk om bogenoemde effek te verwyder.

Ons studie ondersoek of die positiewe korrelasie wat waargeneem word tussen dS en ω 'n artefak van die metodes is wat gebruik word om die beraamde waardes van dN en dS te bepaal. Verder ondersoek ons of die verskeie model-parameteriserings 'n uitwerking het op hierdie korrelasie. Om vas te stel of die korrelasie tussen dS en ω die gevolg is van 'n biologiese eienskap, wat sal beteken dat verdere navorsing nodig is met betrekking tot die verhoudings tussen evolusionêre tempos, en of die korrelasie die gevolg is van 'n statistiese artefak, in welke geval navorsers bedag moet wees op die implikasies van die gebruik van sekere metodes, kan 'n belangrike bevinding wees vir navorsers wat poog om genoomgedeeltes onder positiewe seleksie te identifiseer.

Ons het 'n konstante ω model en 'n punt-tot-punt ω variasie model gepas op die data stel van Drummond et al (2008). Ons het die uitwerking van verskillende kodon-frekwensie beramingsmetodes ondersoek, asook verskillende metodes om die waarde van dN en dS te bereken, op die korrelasie tussen dS en ω bepaal. Ons bevinding is dat die positiewe korrelasie wat waargeneem word tussen dS en ω beïnvloed word deur model-parameterisering. Inderdaad, veranderings in die manier waarop sinverwante en nie-sinverwante posisies gedefinieer word, en die metode waarvolgens die ekwilibrium kodon frekwensies bereken word, het nie net die positiewe korrelasie tussen dS en ω verminder nie, maar in etlike gevalle het hierdie korrelasie negatief geword. Toelating vir punt-tot-punt ω variasie binne 'n geen het ook 'n uitwerking gehad op die korrelasie tussen dS en ω in die empiriese data.

Ons resultate en bevindings brei uit op vorige navorsing wat getoon het dat verskillende modelle van evolusie die positiewe korrelasie tussen dS en ω affekteer. Ons het aangedui watter spesifieke model-parameteriserings en beramingsmetodes die oorsaak kan wees van ongelyksoortige korrelasiepatrone tussen modelle. Ons studie dui aan dat die positiewe korrelasie tussen dS en ω moontlik artifaktueel van aard is en dat hierdie korrelasie dus nie as biologies relevant gesien behoort te word totdat die verhouding tussen dN en dS beter verstaan word nie.

Aknowledgements

I would like to sincerely aknowledge the following for their part in this thesis:

- Firstly, my mom for her support and praise through "write-up week" and so doing adding yet another degree under her belt.
- My father for being the embodiement of hard work and instilling some of it in me.
- My siblings for the fun times in between.
- My supervisor, "Doc", for always demanding perfection and showing me the importance thought.
- Lastly, my beautiful, intelligent wife, Elsamari. For never demanding anything. And always providing everything.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Contextualisation and motivation of the study

In *The Origin of Species*, Darwin stated that a clear understanding of the factors driving the adaptation of organisms to their environment is paramount in the study of evolution [Darwin, 1859]. Darwin and Wallace [1858] theorized that positive selection could explain the process, an idea that remains the episteme of evolutionary theory today. A century and a half later, vast amounts of genome data are becoming available through successful sequencing projects such as the Human Genome Project and the rapid advances in genome sequencing technologies which go hand in hand with such large scale projects [Collins et al., 2003]. These data sets, paired with the the computational capabilities of the modern computer and the development of powerful probabilistic models of evolution, have enabled us to start gaining an understanding of the complex means by which adaptive evolution takes place.

When an organism faces a challenge, be it a competitor, predator or a change in environmental conditions, such as temperature or light intensity, the best solution is often offered by a change in phenotype. A phenotypic change is brought about by a mutation in the DNA of the organism, usually the gene (or network of genes) responsible for the expression of the particular phenotypic trait that needs

to be altered. Depending on where in the DNA the mutation occurs, there is a possibility for the mutation to become a heritable trait to all the future offspring of the organism and hence a probability exists that the mutation will become fixed within a population. When a mutation does become fixed in a population, it is known as a substitution.

Substitutions usually occur due to genetic drift, which is when a mutation becomes fixed in a population by chance. However, if the mutated gene provides the organism with a better-than-average chance of contributing offspring to the population, and hence becomes fixed in the population at a faster rate than would be expected under genetic drift, it is said that the gene is under positive selection [Schaffner and Sabeti, 2008]. Mutations which have the opposite effect on a gene, i.e. a decrease in frequency in a population due to lower fitness, cause the gene to be under purifying selection. A mutation that does not affect the fitness of a gene is neutral [Hellmann et al., 2003]. However, most genes are functionally conserved [Dean and Golding, 1998] and as such are subject to purifying selection. Areas of the genome that are under positive selection are of particular interest as they are often associated with the rapid evolution of pathogens, such as HIV, to escape host immune responses and drug treatments. As such, one of the most frequent uses of probabilistic models of evolution is to detect positive selection acting within a genome.

To detect positive selection, probabilistic models of evolution have been applied to codon data. A codon is a sequence of three nucleotide bases which codes for a certain amino acid when the gene is translated to a protein. The usual methodology to detect positive selection is to compare genomic sequences, for example two homologues (genes which share a common ancestor) from two similar species, and to consider the number of nonsynonymous and synonymous differences between the homologues. A synonymous difference is when the codon mutation between the homologues does not change the amino acid that would be produced by the different homologues during the translation process. A nonsynonymous difference is when a codon mutation between the homologues changes the amino acid that would be produced during the translation process. We can estimate the rate at which nonsynonymous and synonymous substitutions take place in

the data. More specifically, we can estimate the rate of nonsynonymous substitutions per nonsynonymous site (dN) and the rate of synonymous substitutions per synonymous site (dS). These concepts are discussed in depth in Section 2.4.2.

Having obtained the above estimates, one of the most prominent methods used to identify regions in the genome under positive selection involves the comparison of dN to dS [Yang et al., 2000]. It is commonly assumed that synonymous mutations do not influence the evolutionary fitness of a gene since they do not alter protein function. In other words, synonymous substitutions are not the result of selection acting on a gene but rather the result of synonymous mutations becoming fixed within a population through genetic drift. Hence it is assumed that dS is the substitution rate under neutrality. The ratio of dN to dS, commonly represented by the parameter ω in models of evolution [Miyata and Yasunaga, 1980], is therefore used as a measure of selection since nonsynonymous substitutions (measured by dN) are influenced by selection acting on a gene. Mathematically, one would expect the correlation between dS and ω to be negative since, for example, an increase in dS should lead to a decrease in $\omega = \frac{dN}{dS}$. However, dN and dS are based on the same inherent mutation rate in the data. Therefore, if selection is not acting on a gene, any increase in dN would be due to an increase in the mutation rate, which would influence dS proportionally so that ω remains constant. If selection is acting on a gene, an increase or decrease in dN could be due to selection, which would affect dN but not dS, and hence ω would change.

An interesting set of results was presented by Wyckoff et al. [2005]. They found that a positive correlation existed between dS and ω in mammalian data, which implied that dS was positively correlated to selection acting on a gene, a direct violation of the assumption that synonymous substitutions are neutral. A subsequent study by Drummond and Wilke [2008] reaffirmed the positive correlation between ω and dS in five (all except worm) out of six model organisms (human (*H. sapiens*), bacteria (*E. coli*), yeast (*S. cerevisiae*), worm (*C. elegans*), fruit fly (*D. melanogaster*) and mouse (*M. musculus*)).

A closer look at how the parametrisations of the models used in the Wyckoff et al. [2005] and Drummond and Wilke [2008] studies could affect the values of

dN and dS, more specifically the methods used by the authors to calculate dN and dS using the parameter estimates obtained from the models, was warranted. A preliminary literature review revealed that: a) different methods of calculating synonymous and nonsynonymous sites had disparate effects on dN and dS values [Bierne and Eyre-Walker, 2003], b) allowing for ω variation within a gene significantly improved model fit [Yang et al., 2000], c) many methods of equilibrium codon frequency estimation exist (for a review see Rodrigue et al. [2008]), which vary in their accuracy of incorporating gene-specific or species-specific sequence composition properties [Lindsay et al., 2008]. One such property could be codon usage bias (CUB), the term commonly used to describe the observation that genomes contain unequal frequencies of synonymous codons [Plotkin and Kudla, 2010]. CUB might point to selection acting on synonymous mutations (see Duret [2002] for a review), which could influence the correlation between $\omega = \frac{dN}{dS}$ and dS.

In light of the above, we embarked on a study to determine whether these model parametrisation differences, as well as differences in the methods used to calculate dN and dS, could explain the positive correlation between ω and dS discovered by Wyckoff et al. [2005]. We formulated our investigation into the following research objectives.

1.2 Research Objectives

The following objectives were formulated:

- To determine whether the positive correlation between ω and dS reported by Drummond and Wilke [2008] and Wyckoff et al. [2005] was due to the method used to estimate the proportion of synonymous and nonsynonymous sites in the calculation of dN and dS.

- To determine whether the positive correlation between ω and dS reported by Drummond and Wilke [2008] and Wyckoff et al. [2005] could be explained by allowing ω to vary among sites within a gene.

- To determine whether the positive correlation between ω and dS reported by Drummond and Wilke [2008] and Wyckoff et al. [2005] was due to certain equilibrium codon frequency estimation methods being unable to model biological traits, in particular codon usage bias.

1.3 Thesis Outline

In Chapter 2, we review existing research in the field of evolutionary modeling. Key concepts are outlined and common model parametrisations are defined. We present a detailed review of the Wyckoff et al. [2005] and Drummond and Wilke [2008] articles due to their role in our study's conceptualisation, as well as the results of Stoletzki and Eyre-Walker [2011] and Li et al. [2009] who suggested possible explanations for the positive correlation between ω and dS.

Chapter 3 reviews the data set on which our analyses were based and the methods implemented to investigate our research objectives. We also present our empirical analysis results here. Our simulation methods and results are discussed Chapter 4.

Finally, Chapter 5 deals with the interpretations and conclusions of our study and their significance to the field of evolutionary modeling. We identify model parametrisation concepts that need to be considered when inferring positive selection acting on genomes and entertain possible limitations of our study.

Chapter 2

Literature Review

2.1 Introduction

As mentioned in Chapter 1, the common use of ω as an indication of positive selection acting upon a gene has come under scrutiny. The results presented by Wyckoff et al. [2005] posed a challenge to accurate interpretation of ω, since it was shown that ω was positively correlated to dS for a range of organisms. This result was in direct violation of the assumption that synonymous substitutions are neutral with regards to the evolutionary fitness of a gene [Yang et al., 2000, Wyckoff et al., 2005, Drummond and Wilke, 2008]. This suggested an investigation into the manner in which adaptive evolution had been detected up until that point.

Researchers have posed probable explanations for the Wyckoff et al. [2005] results [Stoletzki and Eyre-Walker, 2011, Li et al., 2009, Drummond and Wilke, 2008]. Our study was aimed at further exploring plausible explanations. This chapter reviews current literature, focusing on the above mentioned studies. We review common models of evolution, drawing attention to different parametrisations. We also describe the modeling of certain biological traits, such as codon usage bias (CUB).

2.2 Key Concepts

2.2.1 The genetic code: codons and their role in nonsynonymous and synonymous substitutions

As discussed in Chapter 1, dN and dS are based on the number of nonsynonymous and synonymous changes between genomic sequences when codon models of evolution are applied. This section discusses in more detail how a change is classified as being synonymous or nonsynonymous and gives an overview of the genetic code.

AA	Codon(s) which encode them	AA	Codon(s) which encode them
Ala/A	GCU, GCC, GCA, GCG	Leu/L	UUA, UUG, CUU, CUC, CUA, CUG
Arg/R	CGU, CGC, CGA, CGG, AGA, AGG	Lys/K	AAA, AAG
Asn/N	AAU, AAC	Met/M	AUG
Asp/D	GAU, GAC	Phe/F	UUU, UUC
Cys/C	UGU, UGC	Pro/P	CCU, CCC, CCA, CCG
Gln/Q	CAA, CAG	Ser/S	UCU, UCC, UCA, UCG, AGU, AGC
Glu/E	GAA, GAG	Thr/T	ACU, ACC, ACA, ACG
Gly/G	GGU, GGC, GGA, GGG	Trp/W	UGG
His/H	CAU, CAC	Tyr/Y	UAU, UAC
Ile/I	AUU, AUC, AUA	Val/V	GUU, GUC, GUA, GUG
Met/M	AUG	STOP	UAA, UGA, UAG

Table 2.1: The Universal Genetic Code. The table shows each of the 20 Amino Acids (AA) in existence and the codon(s) which encode them in the RNA. The *stop* codons are also indicated.

The genetic code

Codons are triplets of nucleotide bases in a RNA sequence. These codons are translated to amino acids during the production of proteins within a cell. The arrangement of the 4 bases A, C, G and U(T) into a triplet sequence gives

64 possible combinations. The amino acid which a codon encodes depends on the *genetic code* associated to the particular organism. However, the *universal genetic code*, which we discuss below, applies to most organisms. Under the universal genetic code, three codons are *stop codons* (which do not code for any amino acids) and indicate the end of a sequence to be translated, leaving 61 codons. Only twenty amino acids exist, however, and as a result some amino acids are encoded by more than one codon. This relationship is illustrated in Table 2.1 for the universal genetic code, which we used for all the organisms in this study.

Nonsynonymous and synonymous substitutions

Suppose codon AAU (which codes for the amino acid Asparagine (Asn)) were to undergo a mutation where nucleotide base U were replaced with C. The new codon, AAC, still encodes the amino acid Asparagine (Asn), as can be seen from the table. This constitutes a *synonymous mutation,* since the amino acid encoded by the codon in this position within the sequence remains unchanged, even though the codon has changed. However, consider the situation where the codon AAU changed to AAA. AAA encodes Lysine (LYS). We now have a *nonsynonymous* mutation, since the amino acid encoded by the codon in this position of the sequence has changed. If a mutation becomes fixed in a population, it is known as a substitution. When any nucleotide substitution at the third codon position in a codon does not change the encoded amino acid, the third position is said to be *fourfold degenerate*. For the opposite, where any change at the third position changes the encoded amino acid (such as UGG), the third position is *nondegenerate*.

2.2.2 Pairwise nucleotide gene alignments

The data set in this study consists of pairwise nucleotide gene alignments. Suppose we have a gene which is present in two organisms that share a common ancestor (known as homologues). If these homologues were separated by a speciation event, they are known as orthologues and they can be compared to de-

termine the variation between them. The number of nonsynonymous and synonymous substitutions and the codon frequencies, for instance, are just some of the quantities that are determined from such an alignment. It is common to align *coding regions*, which are the areas of genes which translate to proteins. The text below illustrates a hypothetical pairwise orthologue alignment of coding regions in human and mouse.

```
human_gene_1001      ATG TCG ACT TTT GAC...
mouse_gene_1001      ATG TCG GCT TTT GAG...
```

These orthologues are aligned by nucleotide bases (A,C,G and T). We can also consider it as a codon alignment since each group of three sequential nucleotides forms a codon (such as ATG, TCG and so on). In our example, the final codon in the sequence has changed from GAC (which codes for Aspartic Acid (ASP)) in the human sequence to GAG (which codes for Glutamic Acid (GLU)) in the mouse sequence. This would be a nonsynonymous change.

2.3 Review of existing literature regarding evolutionary modeling

This section reviews current theory and relevant analytical methods with regards to evolutionary modeling and its development. We will start off by retracing the steps to modern theory, followed by model parametrisation definitions (Section 2.3.5) and finally conclude the section with statistical methods of evaluating model accuracy (Section 2.3.8).

2.3.1 A note on phylogenies

Phylogenies form part of the foundation of evolutionary modeling. A phylogenetic tree is a framework for the evolutionary relationship between two or more species or genes. When computing model likelihoods, the likelihood of any given tree is also considered. This computation becomes complex for large trees, but work by Felsenstein [1981] has enabled these model likelihoods to be determined. Our

study was restricted to pairwise gene alignments, hence our trees are simply defined. The reader is referred to the authoritative work on phylogenies by Felsenstein [2004] for a detailed reading.

2.3.2 The inference of positive selection

As discussed, the measure of the nonsynonymous / synonymous substitution rate ratio, ω [Miyata and Yasunaga, 1980] has become the preferred metric for inferring positive selection. ω, which is defined as $\frac{dN}{dS}$, is estimated as a single parameter in models of evolution. dN is the expected number of nonsynonymous substitutions per nonsynonymous site, whilst dS is the expected number of synonymous substitutions per synonymous site [Yang et al., 2000]. Under the widely accepted assumption that synonymous changes are neutral with regards to selection, dS is proportional to the mutation rate. Nonsynonymous changes are then assumed to be due to selection and dN is then viewed as the rate of substitution [Nei and Kumar, 2000, Yang and Nielsen, 2000]. An $\omega > 1$ is therefore seen as an indication of positive selection acting on a gene [Nei and Kumar, 2000, Kosakovsky Pond and Frost, 2005, Yang et al., 2000]. In contrast, purifying selection is implied when $0 < \omega < 1$ and neutral selection when $\omega = 1$ [Yang and Nielsen, 2000].

Although the definition of nonsynonymous and synonymous substitutions is relatively basic and has been discussed as a key concept in Section 2.2, the definition of a synonymous (nonsynonymous) *site* is complex and plays a pivotal part in our study. This definition will be thoroughly discussed in Section 2.4.1 once the parametrisation of codon models has been reviewed.

The estimation of ω is achieved by optimising statistical models over genome data. This computation is made feasible by assuming that the the evolution between genes follows a continuous-time Markov process.

2.3.3 The role of continuous-time Markov processes in models of evolution

A continuous-time Markov process describes a stochastic process, say Z, where Z_t denotes the state of Z at time $t \geq 0$, which at any given time occupies a state, i, from a discrete state space. The process may remain in this space for an arbitrary amount of time until moving into another state, j. The probability of being in state j at time t given that the process is in state i at time 0 is denoted by $P(Z_t = j \mid Z_0 = i)$. A property of such a Markov process is that if, at any time s with $0 < s < t$, the process occupies state k, the probability of the process going from state k to state j from time s to time t can be calculated with the dependence on the history of the process replaced only by the dependence on the state at time s. In short:

$$P(Z_t = j \mid Z_s = k,\, Z_0 = i) \quad = \quad P(Z_t = j \mid Z_s = k). \tag{2.1}$$

Continuous-time Markov processes have been applied to codon evolution modeling by describing the process of codons changing from one state (for instance CTG) to another state (CTT). This enables the computation of the likelihood of observing the data as one need only consider the change of state from codon i to codon j and not the states the codon was in before time i.

The probability for a codon to change from state i to state j in time T is known as a *transition probability*. The transition probabilities for all codons are defined in a *transition probability matrix* (P). In models of evolution, P for any given time T is then

$$P(T) = e^{QT} = I + QT + \frac{1}{2!}(QT)^2 + \frac{1}{3!}(QT)^3 + \ldots \tag{2.2}$$

where Q is the *instantaneous rate matrix*. The off-diagonal entries of $Q = [q_{ij}]$ are the *instantaneous substitution rates*, where q_{ij} represents the rate at which state i changes to state j [Lio and Goldman, 1998].

We define the diagonal elements of $Q = [q_{ij}]$ such that the rows sum to zero: $q_{ii} = -\sum_{j:\, i \neq j} q_{ij}$. This allows spectral decomposition of the Q matrix, which

enables P to be calculated as a function of the eigenvectors and eigenvalues of the Q matrix (see [Lio and Goldman, 1998] for details). $P_{ij}(T)$ is then the probability that codon i will evolve to codon j in time T.

2.3.4 The development of probabilistic models

The most basic models of evolution consider *nucleotide* differences between sequences when estimating the transition probabilities between states. In other words, the instantaneous rate matrix Q is a 4x4 matrix of the instantaneous substitution rates between the nucleotide base pairs A, T, C and G. One of the earlier formulations, introduced by Kimura [1980] made a distinction between transitions (A$\rightleftarrows$G or C$\rightleftarrows$T) and transversions (any other change) between nucleotides. Felsenstein [1981] proposed a model which incorporated the frequencies of the nucleotide bases within the alignment, denoted as π_A, π_T, π_C and π_G. We present the Q matrix defined by Hasegawa et al. [1985] (HKY85) which essentially incorporates the formulations of these two papers:

$$Q = \begin{bmatrix} . & \beta\pi_T & \beta\pi_C & \alpha\pi_G \\ \beta\pi_A & . & \alpha\pi_C & \beta\pi_G \\ \beta\pi_A & \alpha\pi_T & . & \beta\pi_G \\ \alpha\pi_A & \beta\pi_T & \beta\pi_C & . \end{bmatrix} \tag{2.3}$$

where α represents the rate of transitions, β represents the rate of transversions and π_i the frequency of the nucleotide base i $\in$[A, C, G, T]. The diagonal elements of Q are defined such that the rows sum to zero, as discussed in the previous section. Note that the matrix rows and columns are ordered as A, T, C, G. In other words, $q_{1,4}$ ($\alpha\pi_G$) is the instantaneous substitution rate from A to G. The transition/transversion rate bias, $\kappa = \frac{\alpha}{\beta}$, which forms part of many model parametrisations due to the empirical finding that transitions occur much more frequently than transversions [Brown et al., 1982], can then be calculated from this formulation.

Models which consider *codons* when determining the differences between sequences have gained more prevalence in recent years since codon models are

capable of distinguishing between synonymous and nonsynonymous changes and take into account the genetic code of an organism. The models implemented in our study were all codon models and we present their theory in Section 2.3.5.

2.3.5 Codon model parametrisation

The first codon models were proposed in 1994 by Goldman and Yang [1994] (GY) and Muse and Gaut [1994]. The basic GY model underwent minor revisions and was defined as the $M0$ model by Yang et al. [2000]. This was the basic model in our study as Drummond and Wilke [2008] used this model to obtain their results. It therefore forms the template for discussion in this section.

Recall from Section 2.3.3 that the transition probability matrix (P), which contains the probabilities for a codon in state i to to change to state j in a certain time period, is theoretically defined as $P = e^{Qt}$. Q is the instantaneous substitution rate matrix, the entries of which $[q_{ij}]$ are the instantaneous substitution rates.

The instantaneous substitution rate from codon i to codon j ($i \neq j$) is specified as:

$$q_{ij} = \begin{cases} 0 & \text{if } i \text{ or } j \text{ is a stop codon or } i \text{ to } j \text{ requires } >1 \text{ nucleotide substitution} \\ \pi_j & \text{if } i \rightarrow j \text{ is a synonymous transversion} \\ \pi_j\kappa & \text{if } i \rightarrow j \text{ is a synonymous transition} \\ \pi_j\omega & \text{if } i \rightarrow j \text{ is a nonsynonymous transversion} \\ \pi_j\kappa\omega & \text{if } i \rightarrow j \text{ is a nonsynonymous transition} \end{cases} \quad (2.4)$$

where ω represents the non-synonymous/synonymous rate ratio, κ represents the transition/transversion rate bias, and π_j represents the empirical frequency of codon j.

As discussed in Section 2.3.2, ω is the main deciding factor when inferring selective pressure. In the $M0$ model parametrisation, a single ω is estimated for the entire alignment.

An extension of the $M0$ model is the $M3$ model which allows site-to-site ω variation by means of a number of discrete classes which are pre-defined for ω (see Yang et al. [2000]). We implemented the $M3$ model with three discrete classes in our study to see whether allowing for ω to vary between sites could explain the correlation between ω and dS. The model has five additional parameters as we have ω_0, ω_1 and ω_2 with proportions p_0, p_1 and $p_2 = 1 - p_0 - p_1$ respectively.

The empirical codon frequencies are counted from the sample sequence alignments. When codon frequencies are taken as the actual empirical counts in the alignment, it is known as the *F61* parametrisation. An approximation of this is the *F3x4* method, where codon frequencies are calculated by position specific (with regards to the codon position 1, 2 and 3) nucleotide frequencies in the alignment (see Muse and Gaut [1994], Goldman and Yang [1994]). A detailed discussion on these frequency vector estimations is given in Section 2.5.2 as it forms an important part of the study.

An alternative approach to the standard Q matrix parametrisation is to estimate each q_{ij} from large empirical data sets. When this approach is followed, the model is known as an empirical codon model.

2.3.6 Empirical codon model

Most codon models do not take into account the physicochemical properties of the proteins associated to the different amino acids and do not allow for more than one nucleotide substitution per evolutionary time unit. This leaves a substantial proportion of the evolutionary process not being included in current codon modeling approaches [Delport et al., 2008]. However, if codon exchange rate parameters can be estimated empirically from large data sets, these properties will implicitly be accounted for.

This led to the first empirical codon models by Doron-Faigenboim and Pupko [2007] and Kosiol et al. [2007]. These models include codon *exchangeabilities* which are estimated from large codon sequence alignments. They also allow for non-zero instantaneous substitution rates between codons that differ by more than one nucleotide. Such an approach can include site-specific variation, which

could be extended to lineage-specific variation, and could therefore become widely used in selection studies [Anisimova and Kosiol, 2008].

The Empirical Codon Model (ECM) model as proposed by Kosiol et al. [2007] was one of the first models to combine the information obtained from the genetic code in mechanistic models with the information on physicochemical properties of the associated amino acids from empirical models. The instantaneous substitution rate for codon i to codon j ($i \neq j$) is given by:

$$q_{ij} = \begin{cases} 0 & \text{if } i \text{ or } j \text{ is a stop codon} \\ s^*_{ij}\pi_j\kappa(i,j) & \text{if } i \rightarrow j \text{ is a synonymous change} \\ s^*_{ij}\pi_j\kappa(i,j)\omega & \text{if } i \rightarrow j \text{ is a non-synonymous change} \end{cases} \tag{2.5}$$

where s^*_{ij} are the codon *exchangeabilities*, π_j is the empirical frequency of codon j, $\kappa(i,j)$ accounts for transition/transversion bias and ω represents the nonsynonymous/synonymous rate bias.

The s^*_{ij} exchangeabilities were estimated from DNA alignments of 7332 protein families in the PANDIT database [Whelan et al., 2003, 2006]. These exchangeabilities thus implicitly contain the information regarding amino-acid substitutions within the protein structure. Also note that the model makes no restriction on the number of nucleotide substitutions needed to mutate from codon i to codon j.

$\kappa(i,j)$ is a function of the codons i and j, which takes into account whether the change (or overall change when multiple instantaneous nucleotide changes occur) was a transition or a transversion. Since the model allows for multiple instantaneous nucleotide changes, nine variants of this parameter were implemented which is discussed in the paper.

ω represents the non-synonymous/synonymous rate bias. For the ECM, it cannot be interpreted as a simple rate ratio. This stems from the fact that the average nonsynonymous/synonymous bias is inherently included in the model via the s^*_{ij} exchangeabilities derived from the PANDIT database. To quote Kosiol et al. [2007]: "Estimates obtained from mechanistic models, ω_M, and estimates from

the ECM, ω, cannot be directly compared: ω_M represents the absolute non-synonymous/synonymous rate ratio, whereas ω measures the relative strength of selection with respect to an average level implicit in the PANDIT database." A comparison method is suggested in the paper.

As part of the work for this thesis, we developed software templates to implement the Kosiol et al. [2007] model in the phylogenetic analysis package HYPHY [Kosakovsky Pond et al., 2005]. This was published as part of a subsequent study [Delport et al., 2010].

2.3.7 Counting-based methods and the Yang and Nielsen [2000] model parametrisation

Counting-based methods are the predecessors of modern probabilistic models, and were widely used when computational restrictions burdened the optimisation of the probabilistic models. The reader is referred to Nei and Gojobori [1986] and Ina [1995] for an in depth discussion on these methods. These methods commonly involve three steps: 1) the number of synonymous (S) and nonsynonymous sites (N) are counted; 2) the number of synonymous and nonsynonymous differences are counted; 3) a correction for multiple substitutions at the same site is applied before calculating dN and dS. We applied a counting-based method [Yang and Nielsen, 2000] to the Drummond and Wilke [2008] data set, as it enabled a simple investigation into the effect of taking Codon Usage Bias (CUB) into account when calculating dN and dS on the correlation between dS and ω.

The method used by Yang and Nielsen [2000] to estimate dN and dS is perhaps best illustrated by a summary of the iterative algorithm used (as given in the article):

1. Estimate κ from the fourfold-degenerate and nondegenerate sites under an adaptation of the HKY85-based model using empirical codon frequencies.

2. Count the number of synonymous (S) and nonsynonymous (N) sites using the estimated κ and the empirical codon frequencies.

3. Choose starting values for ω and t (sequence divergence level or branch length).

4. Count the number of synonymous and nonsynonymous differences (both transitions and transversions) using κ, the codon frequencies and the current values of t and ω. The transition probabilities are estimated and the proportion of transitional (T) and transversional (V) differences for each synonymous and nonsynonymous site class are calculated.

5. Correct for multiple substitutions to calculate dN and dS using site counts and differences in base frequencies at synonymous and nonsynonymous sites. t and ω are updated using these values of dN and dS.

6. Repeat steps 4 and 5 until the algorithm converges.

The fact that counting methods make no assumptions regarding rate distributions across sites and are computationally faster than probabilistic methods makes them very attractive for large, heterogeneous data sets. However, these methods lack power to detect positively selected sites in small or homogeneous data sets [Kosakovsky Pond and Frost, 2005]. The need for statistically and computationally advanced methods subsequently resulted in the development of probabilistic models.

2.3.8 Hypothesis testing for model selection

Hypothesis tests are used to determine the most appropriate model to be used for a particular biological question. Simply taking the model with the highest maximum likelihood may lead to choosing a model that is unnecessarily complex [Felsenstein, 2004]. The most common hypothesis test for model validity is the Likelihood Ratio Test (LRT). A LRT can be used to compare the maximum likelihood of a simple model, for argument's sake say a model L_0, which does not allow for positive selection, to a more complex model, say L_1, which does allow for positive selection, and determine whether the simple model is inadequate to

explain observations from the data [Felsenstein, 2004]. Under L_0 , the LRT test statistic and its ditribution is given by:

$$2\ln(L_1 - L_0) \sim \chi^2(p) \tag{2.6}$$

where p is the number of extra free parameters that model L_1 has. Rejecting the null hypothesis will indicate the inability of model L_0 to adequately model the data, and that the inclusion of the extra parameters in L_1 is warranted. However, a LRT makes no correction for multiple testing, and for the comparison of non-nested models, the Akaike Information Criterion (AIC$_c$) [Akaike, 1973], corrected for sample size [Sugiura, 1978], is recommended [Anisimova and Kosiol, 2008, Felsenstein, 2004, Delport et al., 2008]. The AIC value for model i is calculated as follows:

$$AIC_i = -2\ln(L_i) + 2p_i \tag{2.7}$$

with p_i equal to the number of parameters. The model with the lowest AIC value is preferred [Felsenstein, 2004].

These tests can be used to select an appropriate model for a sample data alignment since the model with the lowest value is preferred. If a model cannot be chosen with certainty, a model averaging approach can be implemented (as in Kosakovsky Pond and Frost [2005]) but attention needs to be given to parameter interpretation [Anisimova and Kosiol, 2008].

The discipline of evolutionary modeling using codon alignments is undergoing its own rapid evolution. We have reviewed the basic definitions underlying the most commonly used methods at present. We now present a few methodological approaches which are still under debate.

2.4 Methods of calculating dN and dS from model parameter estimates

Recall that dS is the expected number of synonymous substitutions *per synonymous site*, and dN the number of nonsynonymous substitutions *per nonsynony-*

mous site. How the proportion of sites that are synonymous or nonsynonymous within a gene alignment is determined is therefore important and we discuss different methods here.

The second topic we deal with is the actual calculation of dN and dS. Since ω, which represents the nonsynonymous-synonymous rate ratio ($\frac{dN}{dS}$), is a single parameter in models of evolution the actual values of dN and dS need to be calculated separately once MLEs of model parameters have been obtained.

2.4.1 The Mutational Opportunity and Physical Site approach of counting synonymous and nonsynonymous sites.

When calculating the number of synonymous and nonsynonymous sites, one needs first to decide on the appropriate definition of a site. One approach is the Physical Site (PS) definition, where the number of nonsynonymous and synonymous sites are determined from the observed codons. Another approach is the Mutational Opportunity (MO) approach, where sites are determined from Maximum Likelihood Estimates (MLEs) of parameters under the model [Bierne and Eyre-Walker, 2003].

Physical Site

Under the universal genetic code, 25.5% of nucleotide mutations are synonymous, while 74.5% of nucleotide mutations are nonsynonymous [Yang, 2004]. This is due to the fact that all nucleotide changes at the second codon position and most at the first position are nonsynonymous. Furthermore, only about half of the nucleotide changes at the third codon position are synonymous. When following the PS approach, the number of synonymous changes (S_i) and nonsynonymous changes (N_i) of each codon is calculated for all single nucleotide mutations that do not lead to a stop codon ($T_i = S_i + N_i$). See the table below for an example.

Codon	Amino-Acid	Nonsynonymous / synonymous change
CGA	Arg	original
CGT	Arg	S
CGG	Arg	S
CGC	Arg	S
CAA	Glu	N
CCA	Pro	N
CTA	Leu	N
AGA	Arg	S
GGA	Gly	N
TGA	Stop	Stop

Consider the codon CGA, which translates to the amino acid Arg, and the effect of single nucleotide mutations on the amino acid encoded. It can be seen from the table that one of the nucleotide changes leads to a stop codon. We consider the codon as having only 8 possible changes. Of these, 4 are synonymous and 4 are nonsynonymous. We now have that $S_{CGA} = 4$, $N_{CGA} = 4$ and $T_{CGA} = 8$. These quantities are calculated based only on the genetic code and do not take into account the codons present in the alignment.

We now compute the expected number of synonymous sites (S), nonsynonymous sites (N) and total sites (T) for the particular alignment by averaging over the equilibrium codon frequencies:

$$S = \sum_i \pi_i S_i \qquad N = \sum_i \pi_i N_i \qquad T = \sum_i \pi_i T_i \tag{2.8}$$

Under the PS definition, the extent to how synonymous or nonsynonymous a site is therefore depends on the state of the codon at that site at any particular time.

Mutational Opportunity

To obtain the number of synonymous and nonsynonymous sites under the mutational opportunity definition, one needs to calculate the proportion of mutations

that would be synonymous/nonsynonymous under the current evolutionary model [Bierne and Eyre-Walker, 2003]. This is done with ω set to 1 [Yang, 2004] in order for the fixation probability of nonsynonymous mutations to be equal to that of the synonymous mutations. To obtain the proportion of mutations which would be synonymous and nonsynonymous under the current model, and hence the number of synonymous and nonsynonymous sites under the MO approach, we use the following equations:

$$S = \sum_{i,j:\, i \neq j \text{ and } aa_i = aa_j} \pi_i q_{ij} \tag{2.9}$$

and

$$N = \sum_{i,j:\, i \neq j \text{ and } aa_i \neq aa_j} \pi_i q_{ij} \tag{2.10}$$

with

$$T = S + N \tag{2.11}$$

where aa_i represents the amino acid encoded by codon i. The q_{ij} parameters are the instantaneous substitution rates obtained from the Q matrix as defined in Section 2.3.5. Note that the maximum likelihood estimates of the parameters κ and π_i are used, and that $\omega = 1$ is fixed.

It is of interest to note that this method incorporates the transition/transversion rate bias (κ) when calculating the number of sites. Since most transitions lead to synonymous changes, an elevated κ may lead to an increase in the number of synonymous sites, and hence a decrease in dS, when following the MO approach. In cases where κ is not taken into account when estimating sites, such as the PS approach, the number of synonymous sites might be underestimated and an elevated dS might be observed [Yang, 2004].

2.4.2 Estimating dN and dS

Although ω represents $\frac{dN}{dS}$, we have seen that the parameter ω is estimated as a single parameter during model optimisation (Section 2.3.5). The estimation of dN and dS separately can be done after MLEs of the parameters have been determined. As we have shown in the previous section, different methods of determining the number of synonymous and nonsynonymous sites exist. We now discuss the methods of determining the number of synonymous and nonsynonymous changes and show how these are used to estimate dS, the number of synonymous changes per synonymous site, and dN, the number of nonsynonymous changes per nonsynonymous site.

The expected number of substitutions per codon per unit time is given by

$$E[\text{substitution}] = \sum_{i,j:\, i \neq j} \pi_i q_{ij} \tag{2.12}$$

where the MLEs of the parameters are used. This can be divided into the expected number of synonymous and nonsynonymous substitutions per codon per unit time by considering $E[\text{substitutions}] = E[\text{synonymous}] + E[\text{nonsynonymous}]$, the equations of which are:

$$E[S] = \sum_{i,j:\, i \neq j \text{ and } aa_i = aa_j} \pi_i q_{ij} \tag{2.13}$$

and

$$E[N] = \sum_{i,j:\, i \neq j \text{ and } aa_i \neq aa_j} \pi_i q_{ij} \tag{2.14}$$

where aa_i represents the amino acid encoded by codon i. Note that ω is no longer fixed at one as in Equations 2.9 and 2.10 but that the MLE of the parameter is used.

To obtain dS and dN which are the synonymous and nonsynonymous substitution rate *per site*, we normalise the expected number of substitutions per codon over

the number of sites:

$$dS = E[S] \times \frac{T}{S} \tag{2.15}$$

and

$$dN = E[N] \times \frac{T}{N} \tag{2.16}$$

where T, S and N are the quantities calculated by site estimation methods as discussed in the previous section.

2.5 Methods of obtaining codon frequencies from the data set

In this section we discuss codon frequency estimation methods. As mentioned in Section 2.3.5, two prominent methods are F3x4 and F61. The definition and application of each is discussed in more detail and tied in with codon usage bias (CUB).

2.5.1 Codon frequency estimation methods

From the section above, it is clear that parameter values influence the calculation of dN and dS. Different methods of estimating codon frequencies from the data exist, some of which might be more accurate than others. As part of the our research objectives, we wanted to investigate whether more accurate codon frequency estimation methods had an influence on the correlation between ω and dS.

As was mentioned in Section 2.3.5, two prominent methods of calculating equilibrium codon frequencies in an alignment are the F3x4 and F61 methods. Both methods involve determining the frequencies from the observed codons in the

		Base			
		A	C	G	T
Position	1	0.2	0.3	0.25	0.25
	2	0.1	0.4	0.3	0.2
	3	0.25	0.27	0.23	0.25

Table 2.2: The F3x4 codon frequency estimation matrix. This example illustrates a typical F3x4 matrix used to calculate codon frequencies. The method involves 9 parameters which need to be estimated, since the last base for each position can be calculated from the other base frequencies. For instance, the base T in the first position can be calculated as $1 - 0.2 - 0.3 - 0.25$. A codon frequency, for example ACG, can now be determined as $0.2 \times 0.4 \times 0.23 = 0.0184$.

alignment. F3x4 is an approximation of the codon frequencies based on the nucleotide frequencies in the alignment, whilst F61 is an exact count of the codon frequencies.

The F3x4 method involves approximating codon frequencies by counting the number of nucleotide bases in the first, second and third codon positions across the alignment. Each codon frequency is then determined by multiplying these base frequencies together for each codon position. The method has 9 parameters which need to be optimised, as is illustrated in Table 2.2.

The F61 method determines the actual codon frequencies from the alignment. Care must however be taken that the alignment contains enough information to obtain codon frequencies that are representative of the equilibrium codon frequencies in the organism as some codons could be assigned a zero frequency if the alignment is too short. Recall that 64 possible codons exist (Section 2.1). Under the universal genetic code, three codons are stop codons, which leaves 60 codon frequency parameters that need to be determined (the 61st codon frequency is equal to $1 - \sum_{i=1}^{60} \pi_i$). However, If the method can be warranted by means of model comparison methods such as those discussed in Section 2.3.8, it is able to summarise biological factors which improves model fit [Kosiol et al., 2007]. One such factor could be codon usage bias.

2.5.2 Codon usage bias

Codon usage bias (CUB) is a common description given to the phenomenon that organisms (especially those that are fast growing with large population sizes [Yang and Nielsen, 2008]) tend to prefer certain codons above others, even though they may encode the same amino acid. In other words, codons which translate to the same amino acid do not appear in the genome with equal frequencies. Duret [2000] showed that anti-codon tRNA numbers in *C. Elegans* varied for synonymous codons and that codons for which the anti-codon number was highest were preferred to optimise protein translation. Allowing for codon usage bias (CUB) and transition/transversion rate bias ($\kappa = ts/tv$) within a codon alignment is advocated by many studies [Zhou et al., 2009, Yang and Nielsen, 2008, Aris-Brosou and Bielawski, 2006]. In cases where $\kappa \approx 1$ and there is low CUB, not allowing for these effects can produce similar results to models that do allow for them [Yang, 2004]. However, when these effects are present in an alignment, the rate estimates can differ by 300-500% [Yang and Nielsen, 2000].

As one can deduce, the F3x4 method of estimating codon frequencies in an alignment, as discussed in the previous section, might not accurately model extreme forms of CUB. The F61 is more accurate at picking up inequalities in codon frequencies. Rodrigue et al. [2008], however, argue that other factors besides CUB might inadvertently be modeled when an F61 approach is followed, which could have a confounding influence on results. They suggested a model where each codon has a specific codon preference value assigned to it (in other words 60 free parameters), and the actual codon frequencies are estimated by the nucleotide frequencies in the alignment. This approach can better quantify the effect of CUB, whilst minimising the effect that other confounding factors could have on the model.

A review of the codon usage bias effect by Plotkin and Kudla [2010] pointed out that patterns of CUB existed between species, within the genome of a single species and even within genes. CUB has also been detected in many taxa ranging from bacteria and archaea through yeast, fruit fly and worm to mammals. A hypothesis test by Yang and Nielsen [2008] to determine whether the CUB effect could be explained purely by mutational bias in mammals showed this not to be

the case. They concluded that selection may be acting on mammalian genomes through synonymous substitutions. In view of these findings, we wanted to investigate the effect on the positive correlation between dS and ω when CUB was taken into account when estimating dN and dS.

2.6 Studies on the positive correlation between dS and ω

In this section we discuss four research papers which need to be considered in parallel to our study. The first two, Wyckoff et al. [2005] and Drummond and Wilke [2008], formed the basis of the study, as discussed in Chapter 1. The following papers, Li et al. [2009] and Stoletzki and Eyre-Walker [2011], each presented theories for the results of Wyckoff et al. [2005], and therefore warrant a more detailed review.

2.6.1 The positive correlation between dS and ω [Wyckoff et al., 2005]

The results of this paper initiated a debate over the use of ω as an indicator of positive selection. The positive correlation between ω and dS challenged the idea that ω was an accurate indicator of selection (as discussed in Section 2.3.2). The effect was investigated in human-mouse, rat-mouse and human-rabbit data sets and was shown to be present in all of these sets, indicating that the result might extend to all mammals. They did, however, warn that the results were crude due to underlying assumptions.

Although the strength of the correlation was relatively small ($r \approx 0.1$), the authors argued that due to ω variation within the genome and the inherent stochastic noise in the dN and dS parameters, the true correlation between dS and ω was most likely being underestimated and that the observation was to be considered noteworthy. As the stochastic variance of dN and dS increases in short lineages, the authors argued that analysis on such data sets might be

exceedingly inaccurate and corrupt the actual correlation between ω and dS. It was also argued that CpG dinucleotide content in the data and selection acting on dS were not the cause for the correlation.

2.6.2 A reaffirmation of the positive correlation between dS and ω [Drummond and Wilke, 2008]

Following the results of Wyckoff et al. [2005], Drummond and Wilke analyzed six model organisms (human, mouse, fly, worm, yeast and E. Coli) and again found a positive correlation between ω and dS in 5 of the 6 organisms (all except worm). In the supplementary material, they state that the correlation between dS and ω is most probably an artifact of a non-linear dependence between dS and dN and therefore has no biological relevance. They argue that the dependence between dS and dN is, however, biologically relevant. They provide a hypothesis of selection against mistranslation-induced protein misfolding which was shown to explain the covariance between dN and dS during a simulation study, which then causes the positive correlation between dS and ω to disappear. We do however note that the authors had defined the relationship between dN and dS in such a way that they had built in a positive correlation between dS and ω into their simulations. This was done by defining $dS = \exp[N(-1.5, 0.27)]$ and $dN = dS^2 \times \exp[N(-1, 0.7)]$. As one can see from the second equation, $\frac{dN}{dS} = dS \times exp[N(-1, 0.7)]$, which introduces the positive correlation between dS and ω. We will revisit these methods in Chapter 5, after our results have been presented.

Our study was aimed at determining whether the positive correlation between dS and ω presented by this study and the study by Wyckoff *et al.* were possibly a statistical artifact of the methods used to obtain the results. It seems we were not the only ones interested in the cause of these results and two explanations for the positive correlation between dS and ω have been published since the inception of our study.

2.6.3 Explanation 1: The positive correlation between ω and dS is related to substitution model and evolutionary lineage [?]

Following a suggestion by Liao and Zhang [2006] that the positive correlation between dS and ω might be due to model selection, Li et al. fitted a range of counting-based and probability models to human-mouse, mouse-rat and fugu-tetraodon data sets. They found that the correlation varied substantially between models, and that the positive correlation between dS and ω even disappeared when applying certain models. Moreover, the correlation patterns differed between species.

The study showed that the positive correlation between dS and ω was related to evolutionary lineage as closely related organisms displayed a weak positive correlation (human-mouse; $r^2 = 0.28$) whilst distantly related species displayed a weak negative correlation (fugu-tetraodon; $r^2 = -0.215$). They also suggested that the dependence between dS and ω is complex and the measurement of the dependence between these rates via statistics such as a correlation coefficient may not be adequate.

Although this study pointed out that the positive correlation between dS and ω was dependent on the evolutionary model used, the reason for this was not given. They did, however, suggest that: (1) transition/transversion rate bias, (2) codon usage bias, (3) larger rate estimation disparity between models as substitution rate increases and (4) estimation error and imperfect computation for the algorithms could be possible factors influencing the correlations. The authors concluded that care should be taken during the model selection step and that traditional model fit analysis methods might not be applicable to models of evolution.

2.6.4 Explanation 2: The positive correlation between dS and ω in mammals is due to Runs of Adjacent Substitutions [Stoletzki and Eyre-Walker, 2011]

The most recent hypothesis regarding the positive correlation between dS and ω was presented by Stoletzki and Eyre-Walker after an analysis using three models on mouse-rat, human-macaque and human-chimp data.

They argue that although ω values are statistically biased when species are closely related, it is probably not the main factor. They also show that data sets that are of good quality reduce the correlation. In conclusion they suggest that multiple sequential nucleotide substitutions in a gene could be the cause of the positive correlation between dS and ω, and that these substitution runs could be due to mutative or selective forces.

2.7 Conclusion

We have defined key concepts such as the genetic code to orient the reader to the methods used in our study. We have discussed the different modes of selection acting on genomes and shown that the identification of sites at which $\omega > 1$ is still the most widely used criterion for inferring positive selection. We presented continuous-time Markov process theory and showed how the transition probability matrix (P) for codon state changes can be calculated.

We then defined codon models and described their parameters, drawing attention to different methods of estimating equilibrium codon frequencies and methods of defining the instantaneous substitution rate matrix (Q). In Section 2.3.8 we described a procedure for the selection of the best model.

We discussed the Mutational Opportunity and Physical Site definitions of a site and how these are used in the estimation of the evolutionary rates dN and dS. We also presented different methods of obtaining codon equilibrium frequencies and the different degrees to which each of these may allow for biological factors, such as codon usage bias, in genomic data.

Lastly, we reviewed the results by Wyckoff et al. [2005] and Drummond and Wilke [2008] as these presented the basis for our study, whilst the work by Li et al. [2009] and Stoletzki and Eyre-Walker [2011] were reviewed in detail as their studies were also aimed at investigating the positive correlation between dS and ω.

Chapter 3

Empirical Analysis

This chapter presents the methods we used and results of the empirical data analysis. First, we provide an overview of our data set. We then discuss the methods of estimating dN and dS by presenting the different model parametrisations we fitted to the data and the methods we used to estimate dN and dS from these model parameters. We provide the results of our analysis on the empirical data before discussing model fit comparison.

3.1 Data

The original data set of Drummond and Wilke [2008] was provided to us by the authors. The data consisted of reported dN, dS and dN/dS rates per alignment as well as the actual pairwise alignments for six model organisms: human (*H. sapiens*), bacteria (*E. coli*), baker's yeast (*S. cerevisiae*), worm (*C. elegans*), fruit fly (*D. melanogaster*) and mouse (*M. musculus*). Each pairwise alignment contained a DNA coding region aligned to an orthologue (see Table 3.1). Table 3.1 summarises the number of alignments received for each of the organisms and the average alignment length in base pairs.

Model Species	Orthologue Species	Number of Alignments Analysed	Average Alignment Length (base pairs)
H. sapiens	*C. familiaris*	5838	1573
E. coli	*S. typhimurium*	2729	998
S. cerevisiae	*S. paradoxus*	4580	1544
C. elegans	*C. briggsae*	3999	1073
D. melanogaster	*D. yakuba*	7026	1451
M. musculus	*R. norvegicus*	9005	1530

Table 3.1: Summary of alignment data for each of the six model organisms: human (*H. sapiens*), bacteria (*E. coli*), yeast (*S. cerevisiae*), worm (*C. elegans*), fruit fly (*D. melanogaster*) and mouse (*M. musculus*).

3.2 dN and dS estimation

As discussed in Section 2.4.2, the estimation of dN and dS is done using the maximum likelihood parameters obtained from fitting evolutionary models to the data. We therefore present the basic models that we fitted to the data, stating the methods of obtaining codon frequencies and the methods of estimating dN and dS that we used. We also show which software packages we used and statistics that we applied to the data. A summary of abbreviations and parametrisations for all models fitted, as well as the software suite used to fit the model and subsequently estimate dN and dS for each is given in Table 3.2 on page 33.

3.2.1 Model definitions

To obtain model parameter estimates, we fitted a constant rate model ($M0$ - see Section 2.3.5), that is, a model which has one ω value for the entire alignment. The $M0$ model is the model that Drummond and Wilke [2008] used in their study. We also fitted a site-to-site rate variation model ($M3$), in which ω is allowed to vary between sites within an alignment. As discussed in Section 2.3.5, the $M3$ model has a number of discrete classes for ω. We chose three classes, namely ω_0, ω_1 and ω_2 with proportions p_0, p_1 and $p_2 = 1 - p_0 - p_1$ respectively. In order to compare the correlation between dS and ω between the $M0$ and the $M3$ models, we define $\omega* = p_0.\omega_0 + p_1.\omega_1 + p_2.\omega_2$, and use this $\omega*$ value as our

Model Abbreviation	Implementation	Codon Frequency Estimation	Counting of Sites	Allows for Codon Usage Bias
M0_F3x4_PS	HYPHY	F3x4	PS	No
M0_F3x4_MO	PAML	F3x4	MO	No
M0_F61_PS	HYPHY	F61	PS	Yes (implicit)
M0_F61_MO	PAML	F61	MO	Yes (implicit)
M0_F3x4_GF_PS	HYPHY	F3x4	PS	No
M0_F61_GF_PS	HYPHY	F61	PS	Yes (implicit)
YN00_F3x4_CUB_PS	PAML	F3x4	PS	Yes (explicit)
M3_F3x4_MO	PAML	F3x4	MO	No
M3_F61_MO	PAML	F61	MO	Yes (implicit)

Table 3.2: Model Comparison. Models are compared on basis of:
1) Method of calculating number of synonymous and nonsynonymous sites.
2) Allowing for Codon Usage Bias (CUB), either implicitly by means of empirical codon frequencies or explicitly by incorporating codon frequencies into the dN and dS calculation process.
PS : Physical Site approach to counting synonymous and nonsynonymous sites.
MO : Mutational Opportunity approach to counting synonymous and nonsynonymous sites.
F3x4 : F3x4 codon frequency estimation. Codon frequencies are calculated based on position-specific nucleotide frequencies within the alignment.
F61 : F61 codon frequency estimation. Codon frequencies are empirically calculated.
GF : Indicates that a global codon frequency vector (averaged over all *per-alignment* frequency vectors for each organism) was used in place of a per-alignment vector.
$M0$: Corresponds to the $M0$ model implemented by Yang et al. [2000]. The model has one ω ratio for the entire alignment.
$M3$:Corresponds to the $M3$ model implemented by Yang et al. [2000]. The model has three ω classes $(\omega_0, \omega_1, \omega_2)$ with sites classified by empirical Bayes methods i.e. ML parameter estimates are used to determine posterior probabilities of sites belonging to classes. We used an average ω over sites for correlation analyses.
$YN00$: Corresponds to the method of Yang and Nielsen [2000]. Method incorporates Codon Usage Bias (CUB) and transition/transversion rate bias when calculating dN and dS.

ω in our plots of dS vs ω.

For each of the models above, codon frequencies were estimated either using 9 parameters (the F3x4 method) or 60 parameters (the F61 method), as discussed in Section 2.5.1. To further explore the effect of codon frequency estimation methods on the values of dN and dS and hence the correlation between ω and dS, we created codon frequencies that were averaged over the codon frequencies obtained for each of the alignments for each organism. We called them *global* codon frequencies and denote them by GF in Table 3.2.

To investigate the effect of site definition on the positive correlation between ω and dS, we estimated dN and dS for the $M0$ model parameter estimates using both a Mutational Opportunity (MO) and Physical Site (PS) approach.

Finally, after observing that models which incorporated full empirical codon frequencies (F61) seemed to reduce the positive correlation between ω and dS when compared to models using the F3x4 method to determine equilibrium codon frequencies, we applied the Yang and Nielsen [2000] method of calculating dN and dS (denoted YN00 in our study) to the data. A review of the literature showed that Codon Usage Bias (CUB) (Section 2.5.2) seems to play a significant role in biological processes, a phenomenon that would implicitly be modeled by an F61 parametrisation. CUB would also implicitly be modeled by an F3x4 approach, but it could be less accurate if the bias does not match certain patterns. We therefore applied the YN00 method which, although being an approximate counting-based method, takes CUB, by incorporating empirical codon frequencies, and transition/transversion bias into account when approximating dN and dS to investigate our reasoning.

3.2.2 Statistical and phylogenetic software suites

We used two common phylogenetic analysis suites for fitting codon models of evolution: PAML 4.2 [Yang, 1997, 2007], as used by Drummond and Wilke [2008], and HYPHY [Kosakovsky Pond and Muse, 2005b], which allows user-defined custom models. We used PAML to calculate dN and dS using the MO approach, and HYPHY to calculate dN and dS using the PS approach, except

for the $YN00$ model which was implemented in PAML using a PS approach. The use of HYPHY also enabled us to use the global codon frequencies during model fitting. Both suites use maximum likelihood methods for estimating parameters. Batch Processing and general programming was done using Python (`http://www.python.org`).

The Spearman's correlation coefficient between any two mutation rates was calculated using the R Statistical Package [R Development Core Team, 2009]. To test whether differences in the rate correlations between models were significant ($\alpha = 0.05$; one-tailed), we applied the Fisher r-to-z Transformation [Fisher, 1915] to the correlation coefficients. First, a z value is obtained for each model by applying the Fisher Transformation to the correlation coefficient, where $z = \ln \sqrt{\frac{1+r}{1-r}}$. We then subtract the z value of the second model from the first model to obtain a new z value. A value of $z > 0$ indicates that the correlation in the first set is greater than in the second (and $z < 0$ vice versa), this value being accompanied by a $p - \text{value}$ to indicate whether the difference between the correlations is significant.

3.3 Empirical results

3.3.1 Overview

We fitted a constant rate model ($M0$) and a site-to-site rate variation model ($M3$), with various parametrisations, to data from our six organisms in order to assess whether differences in the correlation between dS and ω existed between models and between methods of estimating dN and dS. We also applied the $YN00$ method of estimating dN and dS. We found that the strength of the positive correlation between dS and ω not only varied amongst the different approaches, but that a negative correlation existed in some cases. Figure 3.1 gives an overview of the empirical results for all models and all organisms.

It seems that the main factor influencing the correlation between dS and ω is the manner in which a site is defined when calculating dN and dS, in other words whether a Physical Site (PS) definition (which was used by Drummond and

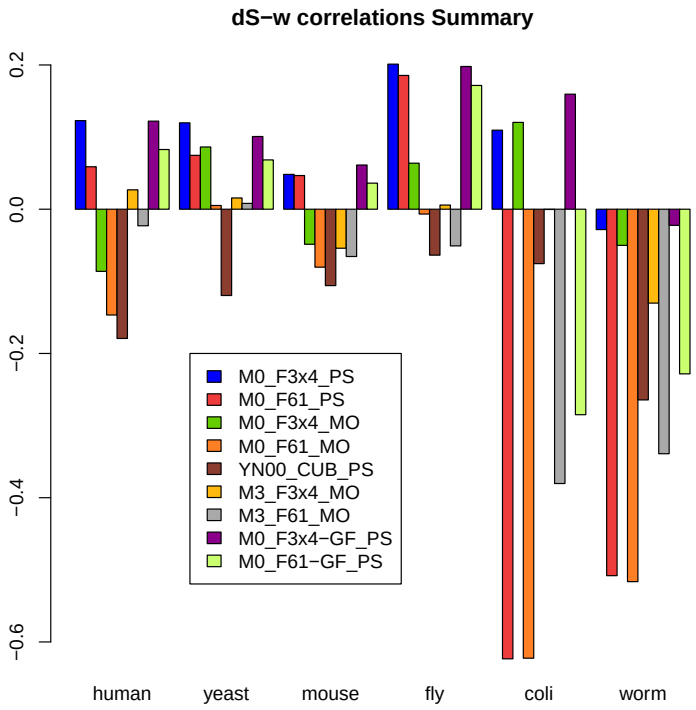

Figure 3.1: A summary of the results we obtained on the empirical data. The figure shows the correlation between dS and ω for the different methods used to estimate dN and dS. The organisms from left to right are human, yeast, mouse, fly, E. coli and worm. Model abbreviations are as defined in Table 3.2 on page 33.

Wilke [2008]) or Mutational Opportunity (MO) definition (see Section 2.4.1) is followed. In two of the five organisms (human and mouse) that showed a positive correlation between dS and ω under the PS definition, the correlation became negative when an MO site definition was used. Two of the remaining three organisms had a 25% (fly) and 70% (yeast) reduction in the correlation using the MO approach when compared to the PS approach.

Another major factor that seems to influence the strength of the correlation is the method used for codon frequency estimation (see Section 2.3.5). In five of the six organisms, the positive correlation between dS and ω was either reduced or became negative when full empirical (F61) codon frequencies were used as opposed to the F3x4 codon frequency estimates (which was used by Drummond and Wilke [2008]).

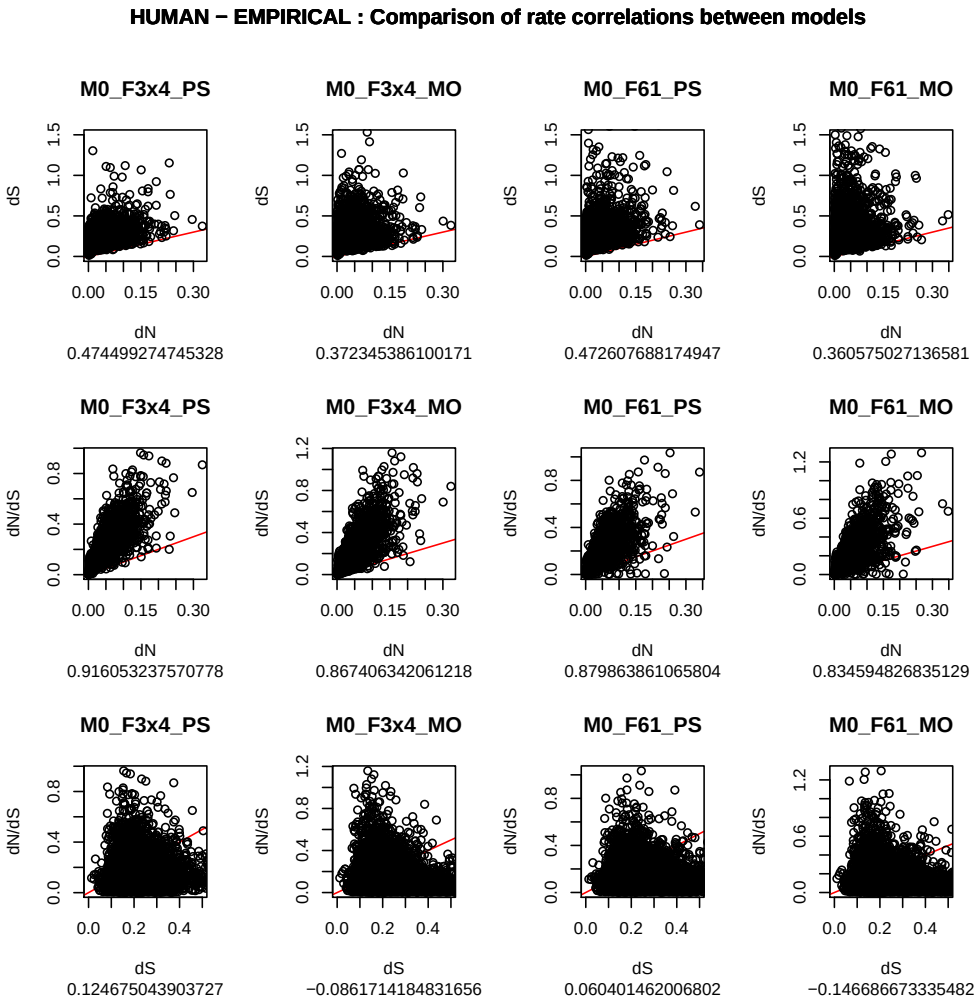

Figure 3.2: The rate correlation pattern as observed in the empirical human data set. We present the correlation between dS and dN, dN/dS and dN, and dN/dS and dS for each of the constant rate model ($M0$) parametrisations. The pattern is similar to those of mouse, yeast and fly. The number below each graph is the Spearman's correlation coefficient. The diagonal line is the $y = x$ line

Finally, the $YN00$ method of estimating dN and dS seems to indicate that Codon Usage Bias (CUB) is a major factor to be considered when estimating rates of evolution as a negative correlation between dS and ω exists for all 6 organisms when this method is applied. Allowing site-to-site ω variation (the $M3$ model)

also reduces the positive correlation between dS and ω for all organisms and even produces negative correlations in some cases.

We found that there were two prominent correlation patterns in the empirical data. The first pattern was similar between human, fly, mouse and yeast and is illustrated in Figure 3.2 by the empirical human data. The pattern has a strong positive correlation between dN and dS, although the strength of the correlation is weaker when using the MO approach. The same observation is reflected in the correlation between dN and ω. Lastly, the positive correlation between dS and ω is either reduced or disappears when MO approaches are applied.

The second pattern was similar for E. coli and worm and is illustrated in Figure 3.3 by the empirical worm data. The pattern also has a strong correlation between dN and dS, but the effect is reduced when models using the $F61$ approach are applied. We also observe a reduced correlation between dN and ω between the F3x4 and F61 models. Lastly, although no notable positive correlation between ω and dS exists in the F3x4 models, a strong negative correlation is observed in the F61 models. We will refer back to these correlation patterns in the simulation analysis chapter. We now present the effect of a) different site definitions b) codon frequency estimation methods c) allowing for CUB and d) allowing ω to vary among sites on the correlation between ω and dS. A full summary of all the results is given in Appendix A.

3.3.2 The effect of different site definitions on the correlation between dS and ω

The use of different site definitions when calculating dN and dS seems to have an effect on the correlation between dS and ω. We present the findings separately for models with an F3x4 (9 parameters) and an F61 (60 parameters) parametrisation.

Figure 3.4 shows the difference in the correlation from the PS approach to the MO approach for the basic $M0$ model with an F3x4 codon frequency estimation approach. The correlation between dS and ω was inverted from 0.1246 to -0.08617 in human and from 0.04827 to -0.04862 in mouse, both correlations differing significantly ($z = 11.44$; $p - \text{value} < 10^{-4}$ and $z = 6.5$; $p - \text{value} < 10^{-4}$

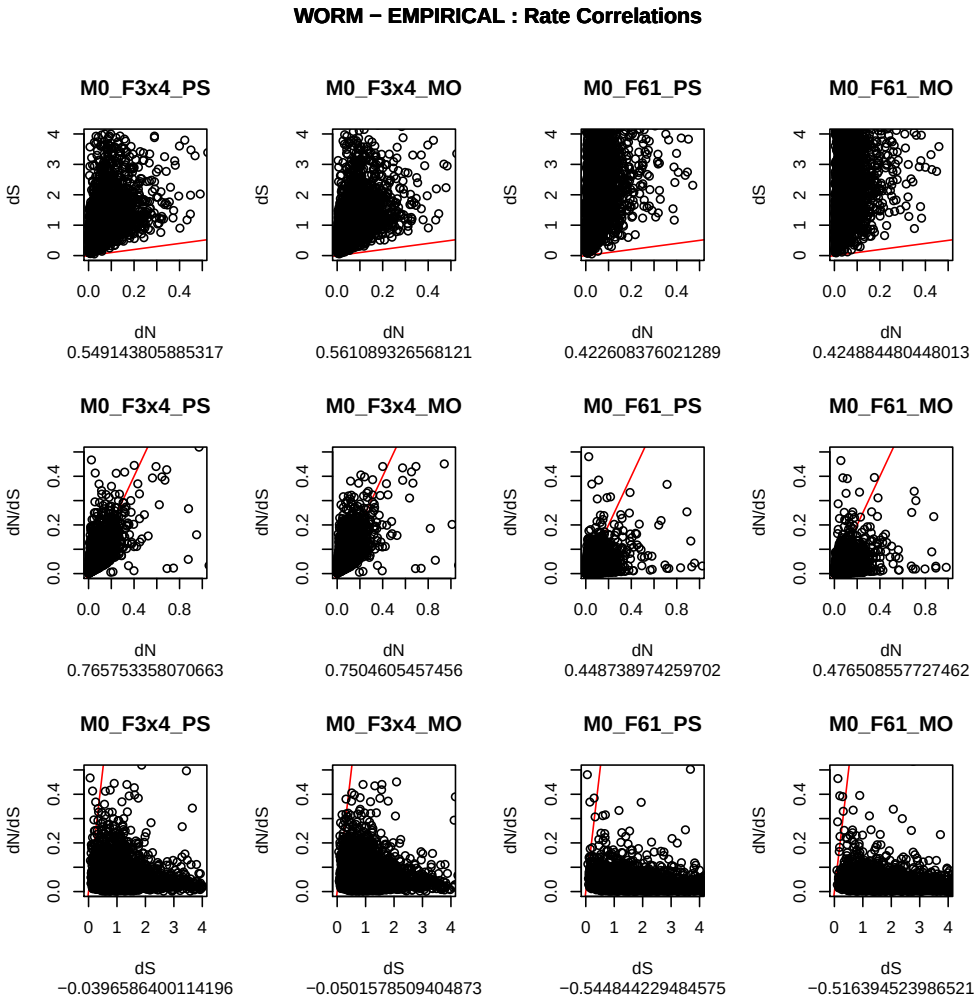

Figure 3.3: The rate correlation pattern as observed in the empirical worm data set. We present the correlation between dS and dN, dN/dS and dN, and dN/dS and dS for each of the constant rate model ($M0$) parametrisations. The pattern is similar to that of E. coli. The number below each graph is the Spearman's correlation coefficient. The diagonal line is the $y = x$ line

respectively). The correlation was reduced from 0.11976 to 0.08628 in yeast, from 0.20115 to 0.0638 in fly and from-0.02822 to -0.05016 in worm. The reduction in yeast and worm was not significant ($z = 1.54$; $p - \text{value} = 0.0618$ and $z = 0.98$; $p - \text{value} = 0.1635$ respectively), but the reduction in fly was ($z = 8.28$; $p - \text{value} < 10^{-4}$). Only in E. coli was there an increase in correlation from 0.10958 to 0.12038, but the result was again not significant ($z = -0.4$; $p - value = 0.3446$)

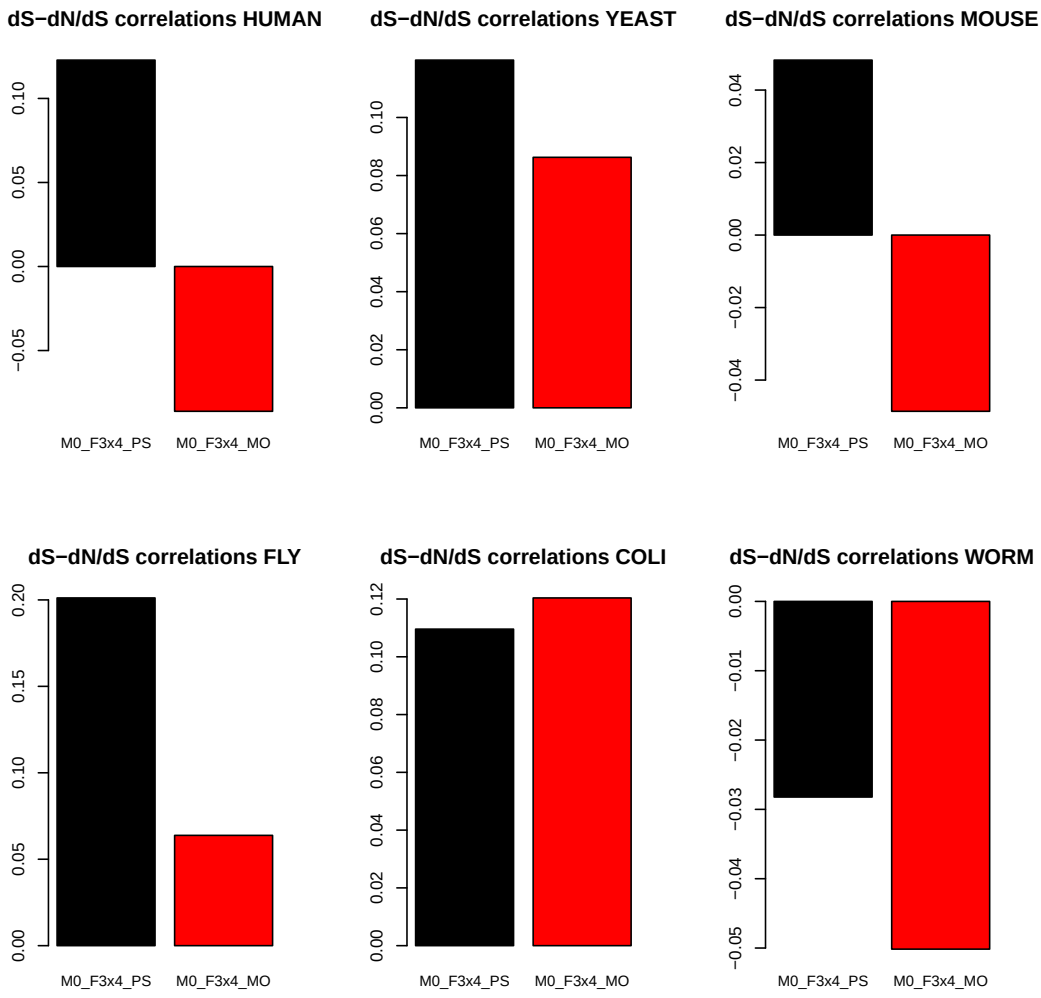

Figure 3.4: The effect of different site definitions on the correlation between dS and ω. We implemented the constant rate model ($M0$) with an F3x4 codon frequency estimation. We present the results for M0_F3x4_PS (black), where the PS approach was used to calculate dN and dS, and M0_F3x4_MO (red), in which the MO approach was used.

The effect of different site estimation methods was very similar in models with an F61 parametrisation, which we present in Figure 3.5. Human, mouse and fly all showed significant inverted correlations between dS and ω from 0.05887 to -0.14669 ($z = 11.13$; $p - \text{value} < 10^{-4}$), 0.0467 to -0.08039 ($z = 8.54$; $p -$

value $< 10^{-4}$) and from 0.18548 to -0.00671 ($z = 11.5$; $p-$ value $< 10^{-4}$) respectively. The correlation practically disappeared in yeast, being reduced from 0.07468 to 0.00517 ($z = 3.3$; $p-$value $= 0.0005$), and remained fairly constant in both E. coli and worm, showing no significant reduction from -0.62341 to -0.62252 and from -0.50812 to -0.51639 respectively.

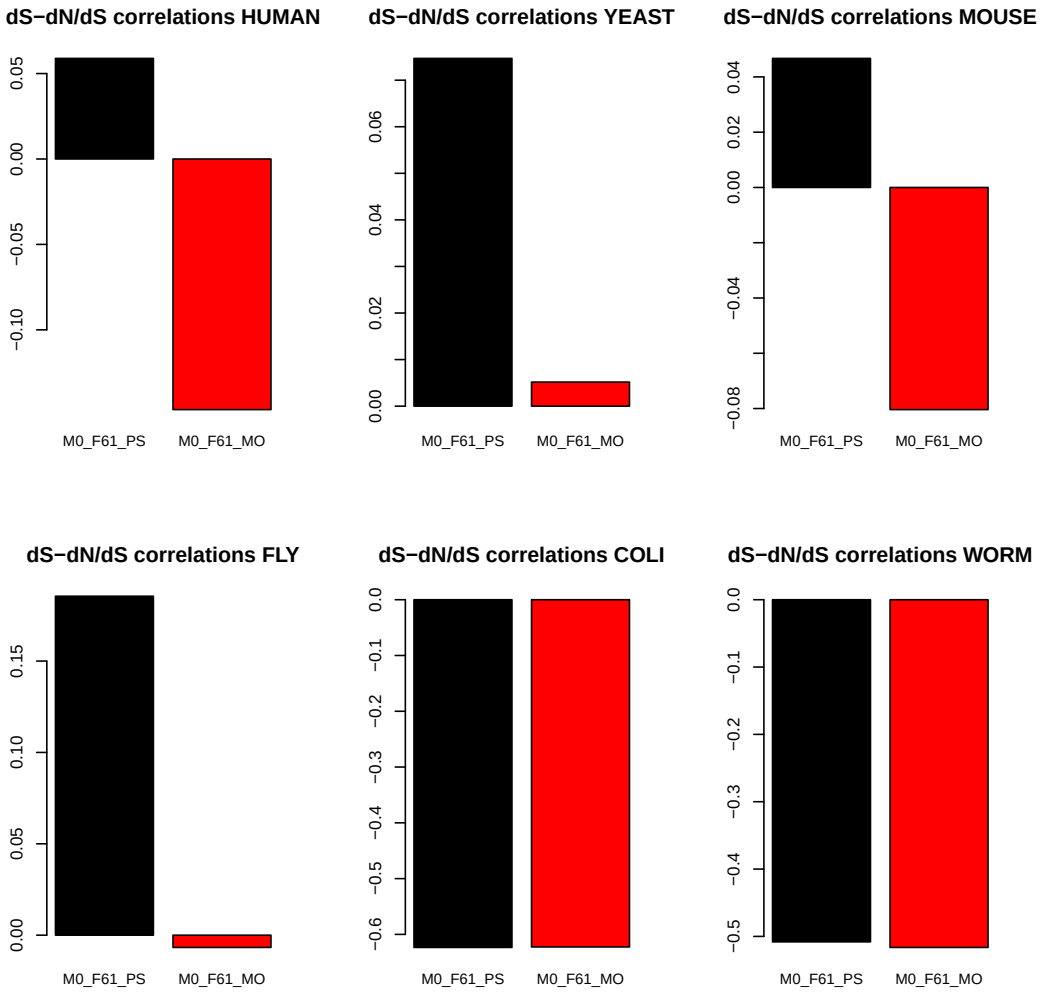

Figure 3.5: The effect of different site definitions on the correlation between dS and ω. We implement the constant rate model ($M0$) with a F61 codon frequency estimation. We present the results for M0_F61_PS (black), where the PS approach was used to calculate dN and dS, and M0_F61_MO (red), in which the MO approach was used.

3.3.3 The effect of different codon frequency estimation methods on the correlation between dS and ω

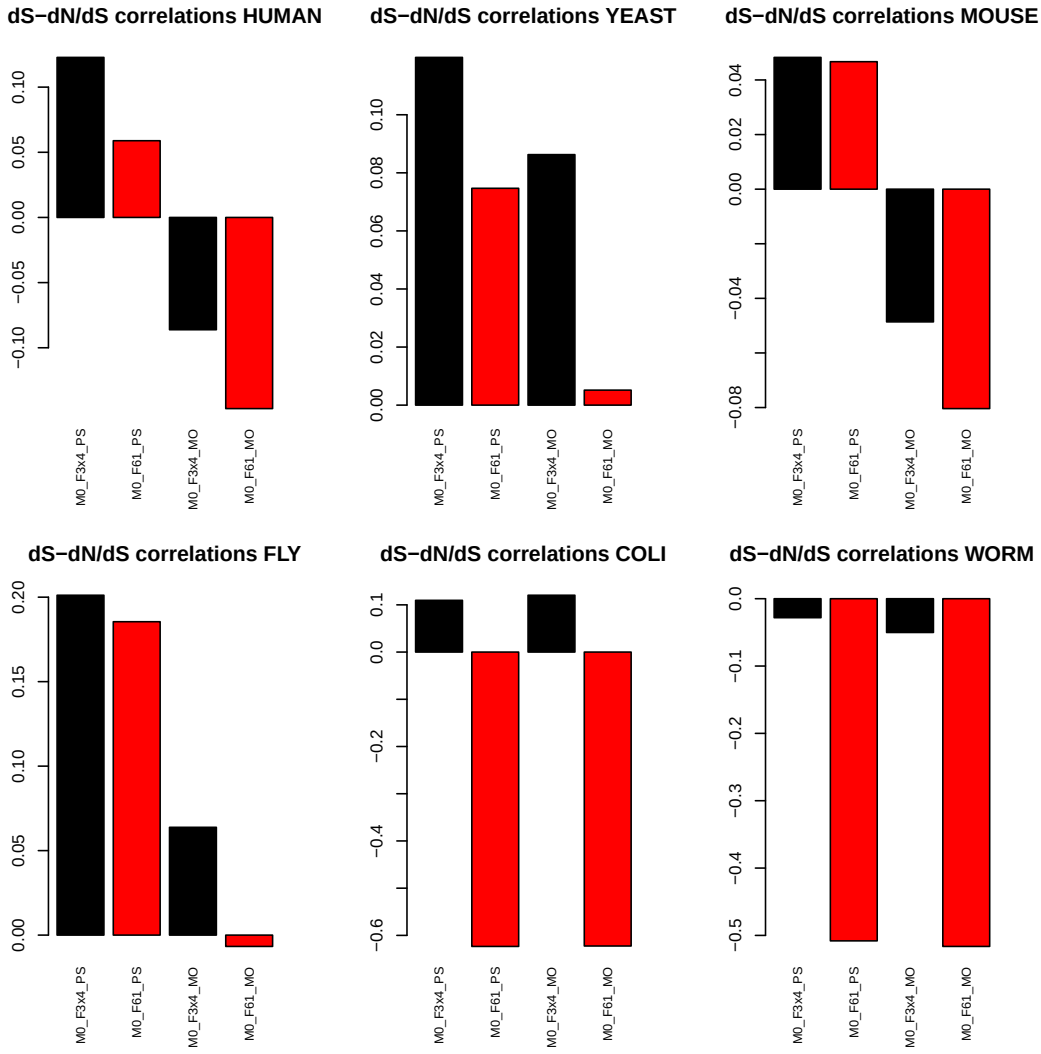

Figure 3.6: The effect of different codon frequency estimation methods on the correlation between dS and ω. The difference in the correlation between dS and ω when using the a F3x4 method (black) and a F61 method (red) is shown. The model pairs, from left to right, are M0_F3x4_PS with M0_F61_PS and M0_F3x4_MO with M0_F61_MO.

Models with different codon frequency estimation methods had markedly different correlations between dS and ω. We present the results for two model pairs which were implemented with both a F3x4 and F61 parametrisation, in other words M0_F3x4_PS with M0_F61_PS and M0_F3x4_MO with M0_F61_MO. The results for models which implemented the global F3x4 and F61 codon frequencies were very similar to those of the M0_F3x4_PS with M0_F61_PS models and therefore we do not present those here.

Figure 3.6 shows that the positive correlation between dS and ω was reduced (or inverted) in every model pairing for all the organisms and therefore we only discuss some of the findings for the sake of readability. In the $M0$_MO models, all six correlation reductions were significant (e.g. human : $z = 3.32$; $p-\text{value} = 0.0005$). In the PS models, human, yeast, E. coli and worm showed statistically significant reductions, whilst the correlations observed in mouse and fly were not significant (mouse: $z = 0.11$; $p-\text{value} = 0.4562$ and fly: $z = 0.96$; $p-\text{value} = 0.1685$). Also of interest is that both E. coli and worm show a significant and strong negative correlation when a F61 approach is followed. E. coli, for example, shows a reduction from 0.10958 to -0.62341 and from 0.12038 to -0.62252 in the PS and MO model pairs respectively.

3.3.4 The effect of incorporating Codon Usage Bias when estimating dN and dS on the correlation between dS and ω

In light of the findings in Section 3.3.3, we wanted to investigate possible properties that could be more accurately summarised by a F61 approach as opposed to a F3x4 approach. The literature review led us to suspect that Codon Usage Bias (CUB) might play a role and we decided to fit the Yang and Nielsen [2000] method (YN00_CUB_PS), which explicitly takes CUB into account.

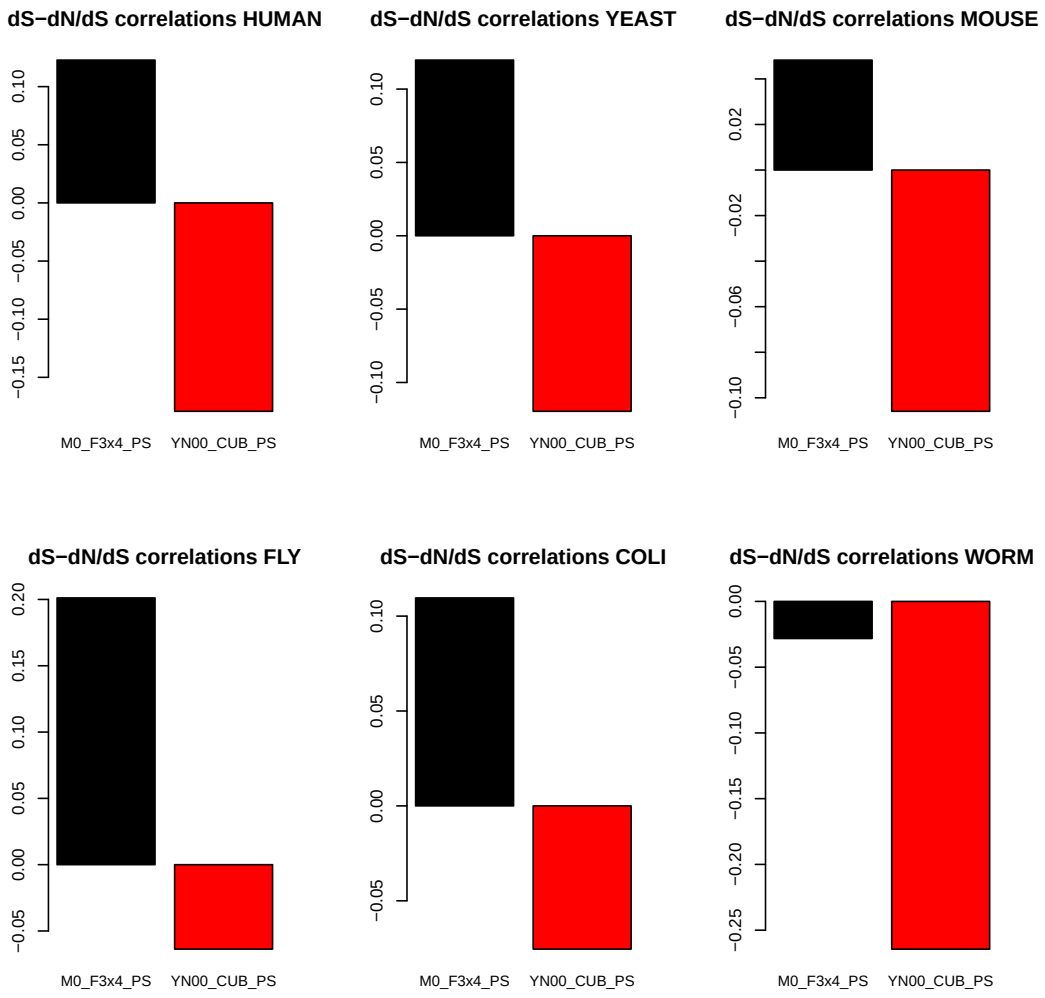

Figure 3.7: The effect of allowing for Codon Usage Bias (CUB) on the correlation between dS and ω. The difference in the correlation between the M0_F3x4_PS model (red) and the YN00_CUB_PS method (black), which allows for CUB when estimating dN and dS, is shown.

The results (see Figure 3.7) show that allowing for CUB when estimating dN and dS has a noteworthy effect on the positive correlation between ω and dS. The correlation became negative in all 6 of the organisms when comparing the $YN00$ method to the $M0$_F3x4_PS method. In the human alignments, we observed a significant inversion from 0.12282 to -0.17922 ($z = 16.47$; $p -$

$\text{value} < 10^{-4}$) and yeast had a similar result with a reduction in correlation from 0.11976 to -0.11966 ($z = 14.23$; $p-\text{value} < 10^{-4}$). Mouse, fly and E. coli had reductions of magnitude 0.1542, 0.26484 and 0.18505 respectively, all of which were significant ($z = 10.37$; $p - \text{value} < 10^{-4}$, $z = 15.83$; $p - \text{value} < 10^{-4}$, $z = 6.85$; $p - \text{value} < 10^{-4}$).

Lastly, although worm had a weak negative correlation between ω and dS to begin with,-0.02822, it became significantly stronger, -0.26442 ($z = 14.35$; $p-\text{value} < 10^{-4}$), using the $YN00$ method. It should be noted that the $YN00$ method follows a PS approach to site estimation, but still showed a statistically significant change from a positive to a negative correlation between dS and ω for all organisms, in contrast to the results in Section 3.3.2.

3.3.5 The effect of allowing for site-to-site ω variation on the correlation between dS and ω

In Figure 3.8 we present the results of the $M0$ model, with dN and dS estimated by both the PS ($M0$_PS) and MO ($M0$_MO) approach as reference against the $M3$ model where dN and dS have been estimated using only the MO ($M3$_MO) approach. As can be seen from the results, even though the correlation differences follow the same pattern as the other model pairs, the correlation between dS and ω does seem to be of a smaller magnitude generally. Note that, as mentioned in 2.3.5, we use the measure $\omega* = p_0.\omega_0 + p_1.\omega_1 + p_2.\omega_2$ as our ω in our plots of dS vs ω.

Once again, a reduction in correlation is seen for all organisms when comparing the $M0$_PS results to $M3$_MO, with mouse and E. coli showing a negative correlation between dS and ω for the F3x4 models. For the F61 models, human, mouse and fly showed a switch from positive to negative correlations, whilst E. coli and worm maintained an already negative correlation albeit that the correlation was now weaker in both species.

A more meaningful comparison is probably gained by comparing the $M0$_MO models to the $M3$_MO models. In all the organisms except human, the F3x4 models once again showed a reduction in the correlation between ω and dS (yeast:

($z = 3.36$; $p - value = 0.0004$, fly: $z = 3.44$; $p - value = 0.0003$, E. coli: $z = 4.66$; $p - value < 10^{-4}$, worm: $z = 3.61$; $p - value = 0.0002$). Human, however, went from a negative correlation, -0.08617, to a positive correlation, 0.02688 ($z = -6.12$; $p - value < 10^{-4}$).

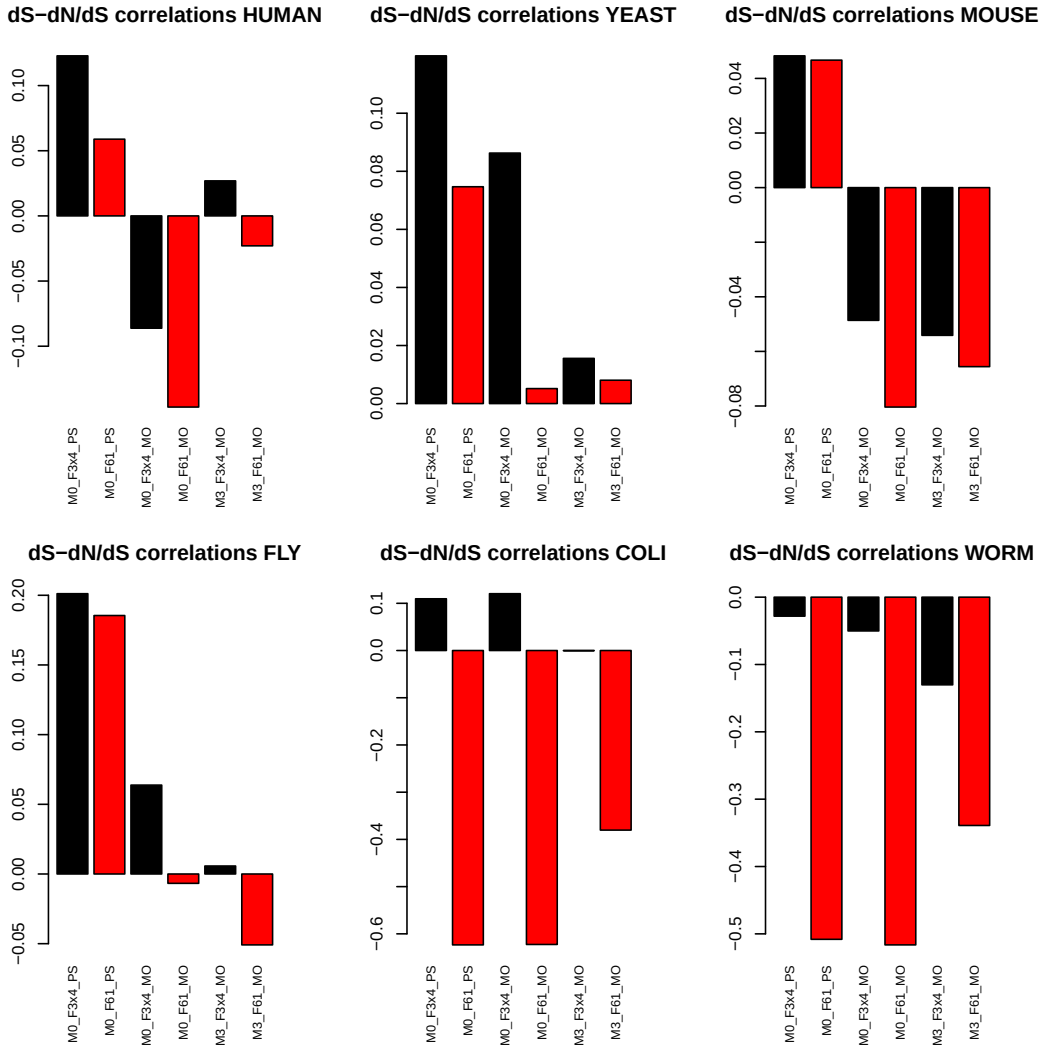

Figure 3.8: The effect of allowing for ω variation on the correlation between dS and ω. From left to right, we present the correlation between dS and ω for the $M0$_PS models, the $M0$_MO models and the $M3$_MO models.

For all the organisms, the $M3$_F61_MO model maintained the same direction of correlation between dS and ω as the $M0$_F61_MO model, but we observed

an overall reduction in the absolute value of the correlations. In other words the reductions became weaker and tended towards a zero correlation between dS and ω. The magnitude of the correlation was significantly stronger for fly ($z = 2.62$; $p-value = 0.0044$), essentially unchanged for yeast ($z = -0.14$; $p-value = 0.44$) and mouse ($z = -1$; $p-value = 0.1587$) and significantly weaker in human ($z = -6.72$; $p - value < 10^{-4}$), E. coli ($z = -12.72$; $p - value < 10^{-4}$) and worm ($z = -9.76$; $p - value < 10^{-4}$).

3.4 Model Fit comparison using AIC criterion

To investigate model fit, we used the AIC criterion as it is most suited for comparing non-nested models (see Section 2.3.8). We were particularly interested to see whether the inclusion of an F61 codon frequency parametrisation was warranted since it had such a marked effect on the results. For each model, we calculated the number of times that the model was within a certain threshold of the AIC value of the best model on each alignment and expressed this as a percentage of the number of alignments in each data set. The difference between the model's AIC and the best model's AIC we denote as $\triangle(AIC)$. If the model itself provided the best fit $\triangle(AIC) = 0$.

The number of parameters for each model were calculated as follows: A constant rate model ($M0$) has 1 parameter for ω, whereas a site-to-site rate variation model ($M3$) has five (ω_0, ω_1, ω_2, p_0 and p_1). A model using the F3x4 codon estimation method has 9 additional parameters, where the F61 method has 60 extra parameters (see Section 2.5.1). When global frequencies are applied, that is the frequencies that are averaged over all the frequencies obtained from the alignments, we do not count the frequencies as parameters that need to be estimated. As shown is Table 3.1, the average base pair length of the pairwise alignments for each organism was as follows: human (1573), E. coli (998), yeast (1544), worm (1073), fly (1451) and mouse (1530).

The results are presented in Table 3.3 and Table 3.4. It can be seen that human and yeast have an overall preference for the $M0$_F61 models, whilst E. coli, worm, fly and mouse all prefer the $M0$_F61-GF models. Since variants of the

Organism	Model	p	$\triangle$(AIC) =0	$\triangle$(AIC)$\leq$ 5	$\triangle$(AIC)$\leq$ 10	$\triangle$(AIC)$\leq$ 20	$\triangle$(AIC)$>$ 20
human	$M0$_F61_PS	62	45.01	5.65	5.36	9.95	33.9
	$M0$_F61-GF_PS	2	31.2	3.61	2.96	5.72	56.42
	$M0$_F3x4_PS	11	10.69	4.06	4.01	8.62	72.54
	$M3$_F61_MO	66	10.52	3.89	6.54	7.88	71.08
	$M3$_F3x4_MO	15	2.02	1.66	4.04	5.45	86.74
	$M0$_F3x4-GF_PS	2	0.46	0.48	1.08	3.65	94.24
yeast	$M0$_F61_PS	62	18.83	4.04	3.97	9.43	63.62
	$M0$_F61-GF_PS	2	61.0	5.2	4.5	5.83	23.35
	$M0$_F3x4_PS	11	11.57	5.18	5.39	11.51	66.24
	$M3$_F61_MO	66	3.69	3.49	8.47	7.25	76.98
	$M3$_F3x4_MO	15	1.57	3.34	6.92	8.89	79.17
	$M0$_F3x4-GF_PS	2	3.21	3.49	4.7	12.86	75.63
mouse	$M0$_F61_PS	62	33.33	5.63	5.81	10.67	44.5
	$M0$_F61-GF_PS	2	48.94	4.22	4.05	6.7	36.01
	$M0$_F3x4_PS	11	9.41	3.42	4.14	9.46	73.51
	$M3$_F61_MO	66	6.22	4.49	11.83	10.88	66.53
	$M3$_F3x4_MO	15	1.53	1.71	5.5	6.73	84.47
	$M0$_F3x4-GF_PS	2	0.52	1.02	1.54	5.95	90.89

Table 3.3: AIC Model Fit Analysis for human, yeast and mouse. p is the number of parameters which need to be estimated. The AIC criterion penalises models based on the number of parameters they contain (see Section 3.2). For each of the models, we show the percentage of alignments in which the model was the best fit ($\triangle \mathrm{AIC} = 0$), and the percentage of alignments which were within a certain range of the best model AIC.

Organism	Model	p	$\triangle$(AIC)=0	$\triangle$(AIC)≤ 5	$\triangle$(AIC)$\leq$ 10	$\triangle$(AIC)$\leq$ 20	$\triangle$(AIC)$>$ 20
fly	$M0$_F61_PS	62	31.66	5.38	4.87	10.01	48.01
	$M0$_F61-GF_PS	2	56.3	4.26	3.54	6.02	29.81
	$M0$_F3x4_PS	11	7.3	2.68	3.7	8.68	77.57
	$M3$_F61_MO	66	3.42	3.49	7.39	7.09	78.55
	$M3$_F3x4_MO	15	0.64	1.11	2.72	5.02	90.43
	$M0$_F3x4-GF_PS	2	0.6	0.93	2.12	7.63	88.64
E. coli	$M0$_F61_PS	62	14.04	3.74	5.06	13.67	63.31
	$M0$_F61-GF_PS	2	61.91	4.11	2.82	6.3	24.67
	$M0$_F3x4_PS	11	1.28	0.73	0.92	2.42	94.46
	$M3$_F61_MO	66	21.66	4.29	5.79	8.94	59.13
	$M3$_F3x4_MO	15	0.44	0.44	0.73	1.8	96.41
	$M0$_F3x4-GF_PS	2	0.44	0.26	0.7	2.02	96.41
worm	$M0$_F61_PS	62	35.85	5.73	5.73	12.86	39.68
	$M0$_F61-GF_PS	2	43.58	4.58	4.3	6.1	41.31
	$M0$_F3x4_PS	11	2.58	1.83	2.1	4.65	88.69
	$M3$_F61_MO	66	16.36	4.45	4.55	7.53	66.98
	$M3$_F3x4_MO	15	0.73	0.48	0.48	1.8	96.4
	$M0$_F3x4-GF_PS	2	0.75	0.68	1.25	3.9	93.29

Table 3.4: AIC Model Fit Analysis for fly, worm and E. coli. p is the number of parameters which need to be estimated. The AIC criterion penalises models based on the number of parameters they contain (see Section 3.2). For each of the models, we show the percentage of alignments in which the model was the best fit ($\triangle \mathrm{AIC} = 0$), and the percentage of alignments which were within a certain range of the best model AIC.

F61 codon frequency estimation method was preferred over the F3x4 method in all of the organisms, we suggest that the reduced correlation between dS and ω which we observed in all organisms when the F61 method was applied might be a truer reflection of the relationship in the data. In other words, the additional information gained by using an F61 method is attractive, even though the method has 60 parameters (compared to the 9 parameters of the F3x4 method). Furthermore, four out of six organisms favoured the global empirical codon frequencies (F61-GF) over the per-alignment empirical frequencies (F61), which could imply that the accuracy gained by using an F61 method on individual alignments is not worth the extra parameters and that the averaged frequencies will suffice for most data sets.

An unexpected result was that the $M0$ model was preferred over the $M3$ model by all organisms, since it has been shown that site-to-site rate variation models provide a better fit to empirical data sets [Kosakovsky Pond and Muse, 2005a]. This could imply that modeling site-to-site rate variation with an extra four parameters is unnecessarily complex and that the constant rate model may provide model parameter estimates that are accurate enough. However, a more likely explanation is that the result is due to the fact that our data consisted of pairwise alignments. There probably is just too little information at each site to accurately apply the $M3$ model to the data. We now present our simulation analysis results, before drawing the empirical analysis and simulations together in Chapter 5.

Chapter 4

Simulation Analysis

4.1 Overview

Drummond and Wilke [2008] stated that the positive correlation observed between dS and ω was probably due to the dependence between dS and dN. This explanation seemed plausible after reviewing the empirical results for the human, mouse, fly and yeast data (see Figure 3.2 for the human results) and the worm and E. coli data (see Figure 3.3 for the worm results). In the human data, we observed a decrease in the positive correlation between dS and dN, and also observed a disappearance in the positive correlation between dS and ω, when comparing the models using the PS approach to the models using the MO approach. The worm data also showed a decrease in the positive correlation between dS and dN, together with a stronger negative correlation between dS and ω, but this time when moving from the F3x4 to the F61 model. Hence, a reduction in the correlation between dN and dS was observed together with a reduction in the correlation between dS and ω when methods were compared, which seemed in line with the findings by Drummond and Wilke.

We wanted to investigate whether different methods caused the reduction in the correlation between dN and dS (and subsequent reduction in the correlation between dS and ω) due to a biological factor that was modeled differently by the methods. As discussed in Section 2.6.2, the authors had defined a non-linear

relationship between dN and dS, and hence had built in a positive correlation between dS and ω into their simulations. It is unclear to us what would cause a non-linear dependence between dS and dN. However, a linear correlation between dS and dN could stem from the fact that both values are based on the inherent mutation rate in the data. For this reason, we simulated data with a positive, linear correlation between dS and dN (Simulations 1-3, Table 4.1). We also wanted to investigate whether a correlation between dS and dN could be introduced by different estimation methods, so we simulated data sets in which no correlation between dS and dN existed (Simulations 4-5, Table 4.1). Lastly, we simulated a data set in which dS and dN were positively correlated, but where ω was varied between sites (Simulation 6, Table 4.1) to investigate what the effect on the correlation between dS and ω would be if a constant rate model ($M0$) was fitted to data in which site-to-site rate variation was present. For all the simulations, we fitted the $M0$ model to the data with both an MO and a PS approach as well as an F3x4 and an F61 approach. In order to observe whether CUB might have an effect on the correlation patterns of the simulated data, the data for Simulations 1 and 4 were generated with equal frequencies for each codon and those for Simulations 2, 3, 5 and 6 with unequal codon frequencies.

4.2 Simulation methods

We simulated data sets under different conditions using the dNdS_Simulator package in the HYPHY suite [Kosakovsky Pond and Muse, 2005b]. The dNdS_Simulator package allows one to specify dS and ω for a number of different sites in each simulated alignment. We were therefore able to specify (dS, ω) pairs for our simulated genes to obtain the desired correlations between dN and dS. The transition/transversion ratio (κ) was set at 3 and the branch length (t) set at 0.6 for all simulations, in line with the observed values for the empirical data.

Our first group of simulated data sets (Simulations 1-3, Table 4.1) was set up in such a way that dN and dS were positively correlated in the generated alignments. For *Simulation 1*, we generated nine subsets of 650 alignments, each alignment being 500 codons in length (total data set: $9 \times 650 = 5850$ alignments).

Each subset of 650 alignments was generated under a different dS - ω pairing, in other words each simulated gene in the subset was generated with the same dS - ω relationship at all sites within the gene. The nine pairings for (dS, ω) were: (0.2, 0.2), (0.2, 0.8), (0.2, 1), (0.6, 0.2), (0.6, 0.8), (0.6, 1), (0.9, 0.2), (0.9, 0.8) and (0.9, 1). By using these values, we should, in theory, have a data set with no correlation between dS and ω and a positive correlation between dS and dN ($r \sim 0.632$). The subsets for Simulation 1 were also generated with equal codon frequencies ($1/61$) across each alignment. *Simulation 2* was set up identically to Simulation 1, except that the data was generated with unequal codon frequencies in each alignment. In *Simulation 3*, we wanted to generate a data set which inherently had a positive correlation between dS and ω. We generated six subsets of 1000 alignments, each alignment 500 codons in length (total data set: 10×600 = 6000 alignments). Each subset again contained genes which were generated with one of the following six (dS, ω) pairings: (0.2, 0.2), (0.2, 0.8), (0.2, 1), (0.6, 0.8), (0.6, 1), (0.9, 1). These pairings have a strong correlation between dS and dN ($r \sim 0.92$) and a positive correlation between dS and ω ($r \sim 0.48$). We used unequal codon frequencies in Simulation 3.

Our second group of simulations were set up in such a way that there was very little correlation between dS and dN (Simulations 4-5, Table 4.1). *Simulation 4* consisted of nine subsets of 650 alignments, each alignment being 500 codons in length (total data set: 9×650 = 5850 alignments). The alignments in each subset were generated with a (dS, ω) pairing chosen from the following: (0.2, 0.02), (0.2, 1.4), (0.2, 0.65), (0.6, 0.9), (0.6, 0.065), (0.9, 0.06). In theory, this data set would have little correlation between dN and dS ($r \sim 0.03$), and a negative correlation between dS and ω ($r \sim -0.216$). Equal codon frequencies were used for the alignment. *Simulation 5* was set up identically to Simulation 4, except that unequal frequencies were used to generate the data.

Our final simulation, *Simulation 6,* was aimed at exploring the effect of allowing dS and ω to vary between sites within each alignment. Three subsets each containing 2000 alignments were generated (total data set was 3× 2000 = 6000 alignments). The alignments in each subset consisted of 498 (3x166) codons, where each group of 166 codons was generated with a different dS-ω pairing.

Simulation #	Correlation between dS and ω	Correlation between dS and dN	ts/tv ratio (κ)	Codon Frequencies	Branch Length (t)
1	0	0.632	3	Equal	0.6
2	0	0.632	3	Unequal	0.6
3	0.48	0.92	3	Unequal	0.6
4	-0.216	0.03	3	Equal	0.6
5	-0.216	0.03	3	Unequal	0.6
6	0	0.767	3	Unequal	0.6

Table 4.1: Simulation Summary. The table shows the correlation between the evolutionary rates under which our alignments were simulated. The table also shows whether the data set in each simulation was generated with unequal or equal codon frequencies.

The dS-ω pairings were: (0.2, 0.2), (0.2, 0.8), (0.2, 1) for group 1, (0.6, 0.2), (0.6, 0.8), (0.6, 1) for group 2 and (0.9, 0.2), (0.9, 0.8) and (0.9, 1) for group 3. There was a strong positive correlation between dS and dN for the data ($r \sim 0.77$), and no correlation between dS and ω. We used unequal codon frequencies in the simulation.

We fitted the $M0$ model with both an F3x4 and an F61 codon frequency estimation approach to each simulation. We then used both the the MO and PS approach to obtain estimates of dN and dS from the simulated data. The simulations are summarised in Table 4.1. It is important to note that the correlations in the table are the actual correlations for the distribution from which the data were simulated, in other words the correlation between the (dS, ω) pairs under which the data were simulated. The correlation between the estimated dS, dN and ω rates which are obtained after fitting a model to the data are reported in the results.

4.3 Simulation results

4.3.1 Simulation 1: Positive correlation between dS and dN and equal codon frequencies

As can be seen from Figure 4.1, even though the data were generated with a positive correlation between dS and dN ($r \sim 0.632$) and no correlation between dS and ω, there is a negative correlation between the estimated values of dS and ω. The negative correlation that we observe in the estimated data is probably due to estimation error, either in the model parameters or in the methods used to calculate the sites. Although a slight difference in the correlation between dS and ω exists between the F3x4 and F61 models, the difference is not significant (e.g. $M0$_F3x4_PS vs $M0$_F61_PS: $z \sim 1.16$; $p - \text{value} = 0.123$).

4.3.2 Simulation 2: Positive correlation between dS and dN and unequal codon frequencies

Figure 4.2 again shows that, even though the data were generated with a positive correlation between dS and dN ($r \sim 0.632$) and no correlation between dS and ω, there is a weak negative correlation between the estimated dS and ω, a result which is again probably due to estimation error as discussed above. Overall, a significantly lower correlation between dS and ω exists for all models between this data set and the data generated with equal frequencies for each codon (Simulation 1) (e.g. $M0$_F3x4_PS (equal) vs $M0$_F3x4_PS (unequal): $z \sim 1.68$; $p-\text{value} = 0.0465$). Also, a significant reduction in the correlation between dS and ω is observed between the F3x4 and F61 models (e.g. $M0$_F3x4_PS vs $M0$_F61_PS : $z \sim 2.32$; $p - \text{value} = 0.0102$). This suggests that methods that are able to accurately model unequal codon frequencies (or CUB) in genes reduce the correlation between dS and ω. Furthermore, the F61 method is richer than the F3x4 method and therefore further significantly reduces the correlation between dS and ω.

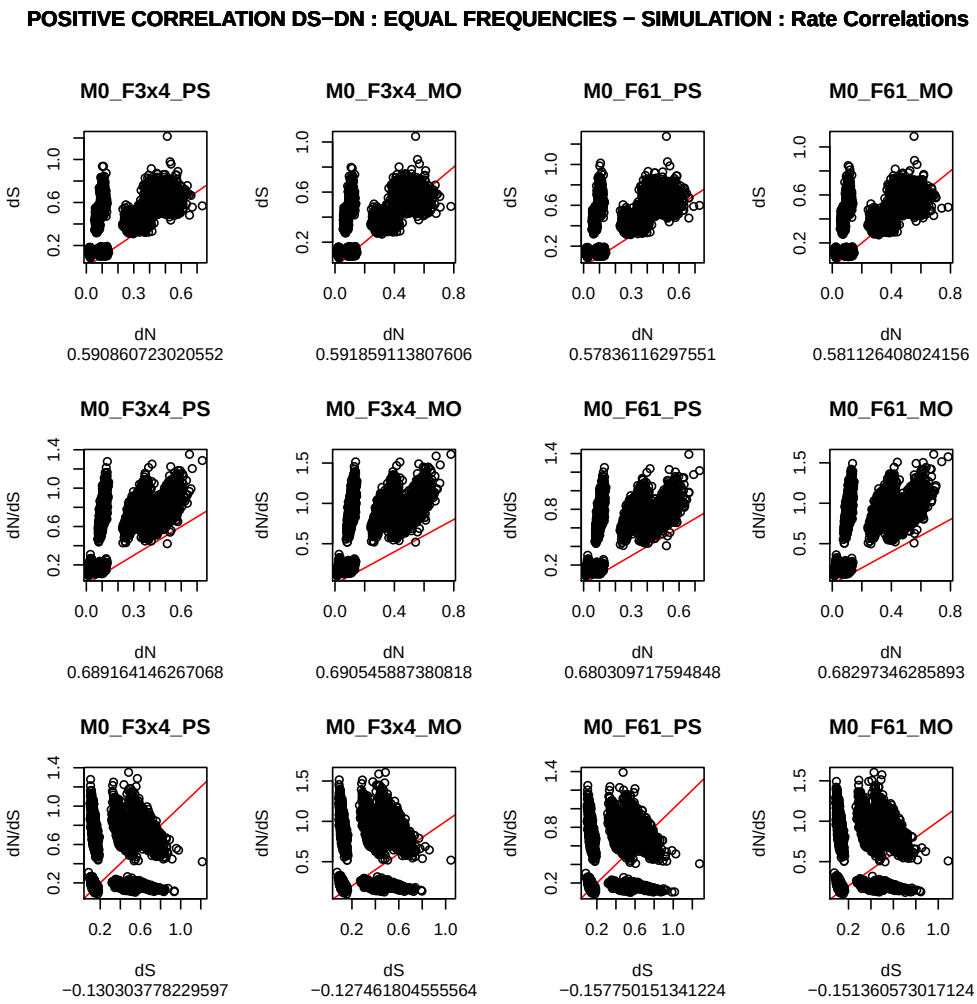

Figure 4.1: Simulation 1: Data set generated under a positive correlation between dN and dS ($r \sim 0.632$) and equal codon frequencies. The correlation between dS and dN, dN/dS and dN, and dN/dS and dS for each of the constant rate model ($M0$) parametrisations is shown.

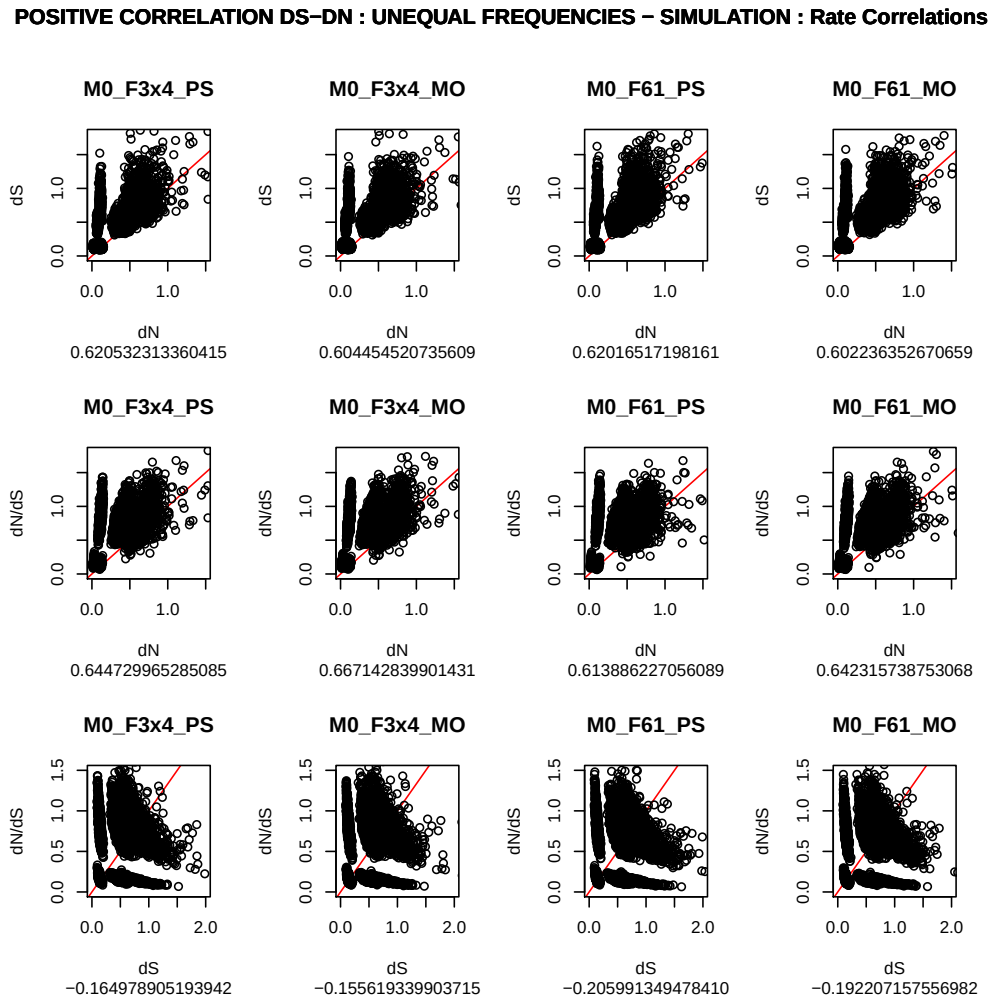

Figure 4.2: Simulation 2: Data set generated under a positive correlation between dN and dS and unequal codon frequencies. The data show the correlation between dS and dN, dN/dS and dN, and dN/dS and dS for each of the constant rate model ($M0$) parametrisations.

4.3.3 Simulation 3: Positive correlation between dS and dN, positive correlation between dS and ω, and unequal codon frequencies

We introduced a positive correlation between ω and dS into the data set, wanting to see whether the models would again show the reduction in correlation between the MO and PS approaches, as well as between the F3x4 and F61 approaches. The results are presented in Figure 4.3. The models show a significant reduction in the correlation between dS and ω between the F3x4 and F61 models (e.g. $M0$_F3x4_MO vs $M0$_F61_MO : $z \sim 2.66; p - \text{value} = 0.0039$), which is consistent with our findings in Simulation 2. Increases in the correlation between dS and ω was observed from the PS to the MO models (e.g. $M0$_F61_PS vs $M0$_F61_MO : $z \sim -1.57; p - \text{value} = 0.0582$), but the results were insignificant.

4.3.4 Simulations 4&5: Zero correlation between dS and dN.

These simulations were aimed at investigating whether the positive correlation between dS and ω would manifest in a data set in which no dependence between dS and dN existed, which could indicate that the effect could be introduced by different methods. The correlation patterns between Simulation 4 and 5 were not dissimilar and we therefore only present the results of Simulation 5 here, in which no correlation between dN and dS existed and codon frequencies were unequal. As is shown in Figure 4.4, there is no correlation between dS and dN and we do not observe a positive correlation between dS and ω. The slight differences observed between the PS and MO models are not significant (e.g. $M0$_F3x4_PS vs $M0$_F3x4_MO: $z \sim 1.34; p - \text{value} = 0.0901$), nor are the differences observed between the F3x4 and F61 models (e.g. $M0$_F3x4_PS vs $M0$_F61_PS: $z \sim 0.29; p - \text{value} = 0.3859$). This could suggest that the cause of the positive correlation observed between dN and dS in the empirical data is modeled to different extents in our group of models, which has an effect

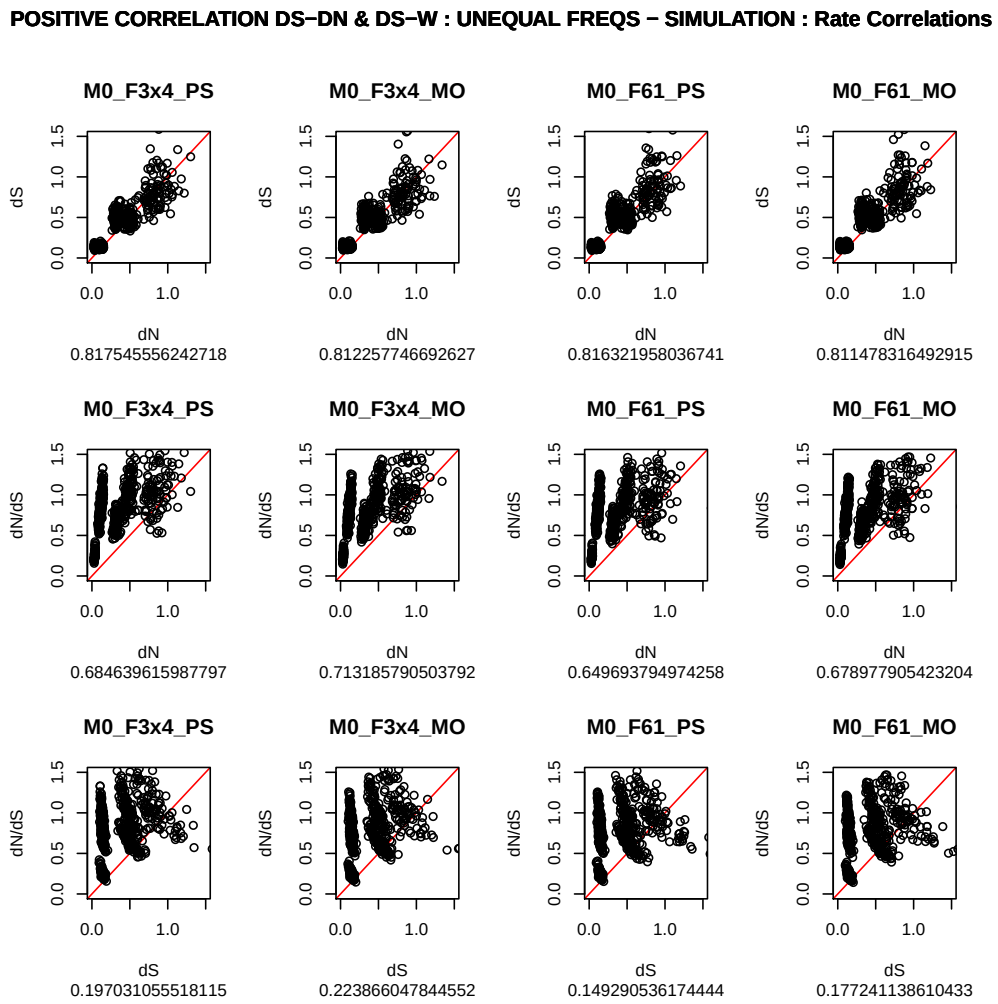

Figure 4.3: Simulation 3: Data set generated under a positive correlation between dN and dS, a positive correlation between ω and dS, and unequal codon frequencies. The data show the correlation between dS and dN, dN/dS and dN, and dN/dS and dS for each of the constant rate model ($M0$) parametrisations.

on the resulting correlations between ω and dS for each model, as we observed in the previous simulations. When the cause is removed, and no positive correlation between dN and dS exists, there is no significant difference in the correlation between dS and ω between models. It also shows that a correlation between dN and dS is not introduced by methods used to estimate these rates.

4.3.5 Simulation 6: ω varied between sites within the alignment

Our final simulation was aimed at investigating the effect on the correlation between ω and dS when ω was varied among sites within the gene. The data were generated with a strong correlation between dN and dS ($r \sim 0.77$) and unequal frequencies. Figure 4.5 shows that, even though a strong positive correlation exists between the estimated dN and dS ($r \sim 0.88$), a strong negative correlation exists between dS and ω ($r \sim 0.82$). The magnitude of the negative correlation is significantly reduced when comparing the PS models to the MO models (e.g. $M0$_F3x4_PS vs $M0$_F3x4_MO: $z = -3.19$; $p - value = 0.007$), an observation which we also made in Simulation 3, where there was a strong positive correlation between dN and dS and a positive correlation between dS and ω. We did however again observe a significant increase in the magnitude of the negative correlation between dS and ω (e.g. $M0$_F3x4_PS vs $M0$_F61_PS: $z = 4.48$; $p - value < 10^{-4}$) for the F3x4 versus the F61 models. This observation seems to be consistent for most of our simulations and empirical data.

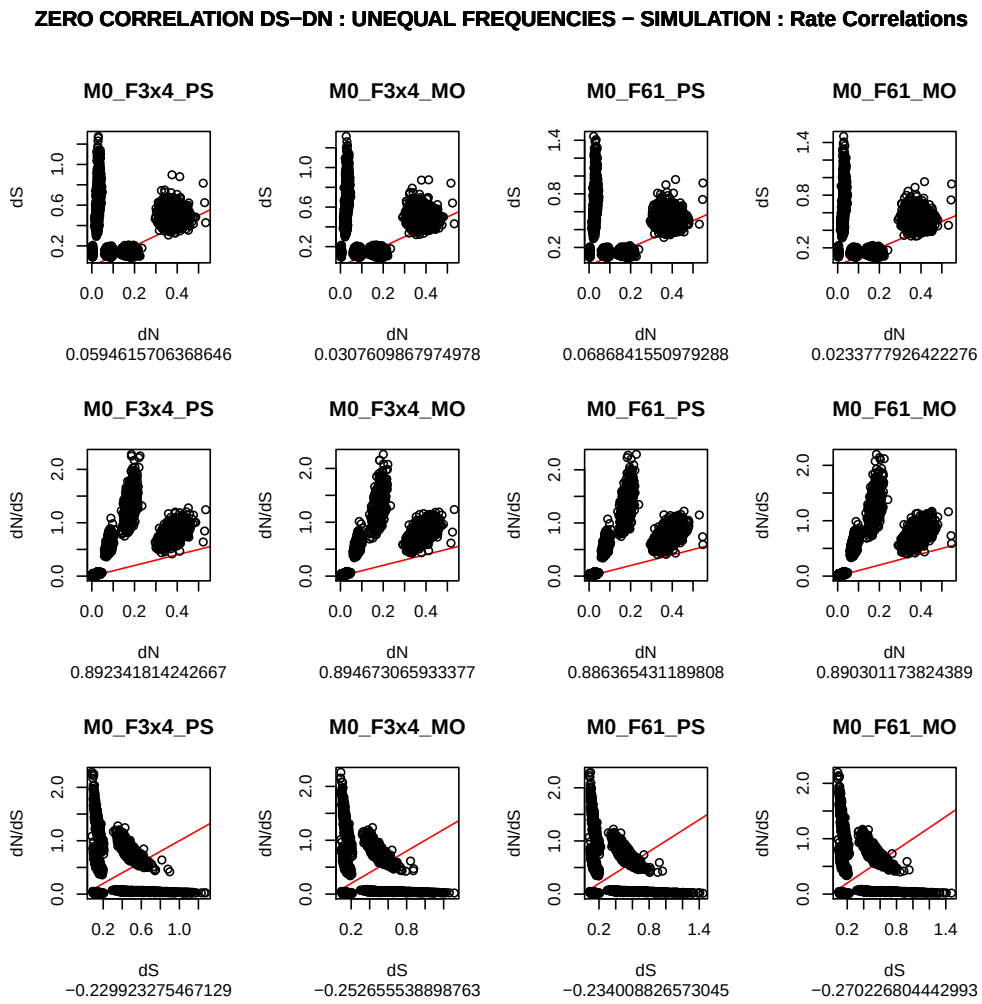

Figure 4.4: Simulation 5: Data set generated under a zero correlation between dN and dS and unequal codon frequencies. The data show the correlation between dS and dN, dN/dS and dN, and dN/dS and dS for each of the constant rate model ($M0$) parametrisations.

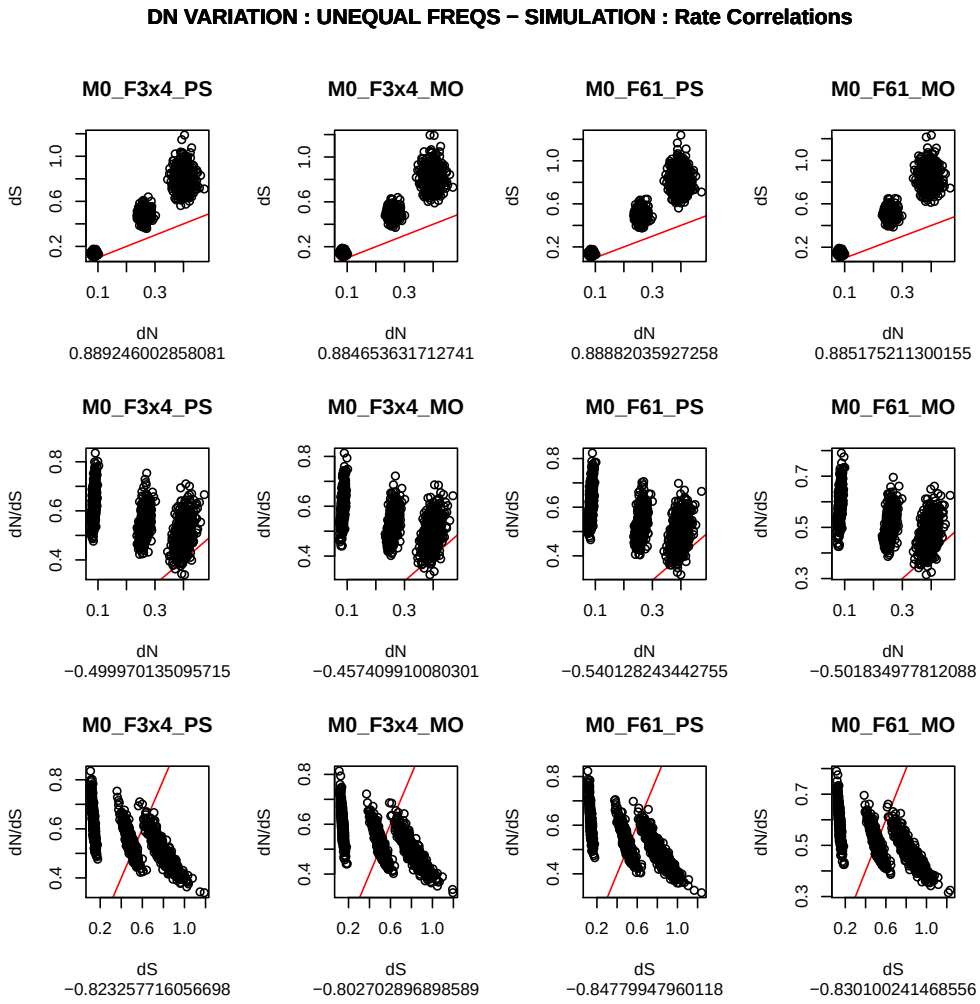

Figure 4.5: Simulation 5: Data set generated with dN varied between sites whilst keeping dS constant. The data were generated with a strong correlation between dN and dS ($r \sim 0.77$) and unequal frequencies. The data show the correlation between dS and dN, dN/dS and dN and dN/dS and dS for each of the constant rate model ($M0$) parametrisations.

Chapter 5

Conclusions and Recommendations

5.1 Introduction

Wyckoff et al. [2005] found an unexpected positive correlation between dS and ω, but could not explain the result. Drummond and Wilke [2008] reaffirmed the result in six model organisms and hypothesised that the correlation could be due to selection acting against protein misfolding. Two further studies on the cause of the positive correlation between dS and ω were conducted during the course of our study. Stoletzki and Eyre-Walker [2011] proposed that multiple sequential nucleotide substitutions could be a cause. Li et al. [2009] fitted a series of models to human, mouse and fugu data sets and showed that the correlation between dS and ω varied considerably between models, and between the evolutionary distance between orthologue species. They also suggested a range of possible factors (see Section 2.6.3).

We formulated three research objectives to investigate the cause of the positive correlation between dS and ω, namely:

- To determine whether the positive correlation between ω and dS reported by Drummond and Wilke [2008] and Wyckoff et al. [2005] was due to the

method used to estimate the proportion of synonymous and nonsynonymous sites in the calculation of dN and dS.

- To determine whether the positive correlation between ω and dS reported by Drummond and Wilke [2008] and Wyckoff et al. [2005] could be explained by allowing ω to vary among sites within a gene.
- To determine whether the positive correlation between ω and dS reported by Drummond and Wilke [2008] and Wyckoff et al. [2005] was due to insufficiencies in the equilibrium codon frequency estimation methods to model biological traits, in particular Codon Usage Bias.

To pursue these research objectives, we fitted a range of models with varying parametrisation options to the Drummond and Wilke [2008] data. We conclude this study by interpreting our empirical and simulation results before discussing possible implications of our findings and recommendations for future research for each of our research objectives separately.

5.2 Interpretation of the empirical and simulation results

The method used to determine nonsynonymous and synonymous sites (PS or MO) had a definite impact on the correlation between dS and ω in the empirical results for human, mouse, fly and worm. The methods are identical in all regards except for the way in which the proportion of synonymous and nonsynonymous sites is estimated. As discussed in Section 2.4.1, the estimation of dN and the estimation dS are both based on the maximum likelihood parameter estimates (κ, π_j etc.), which could introduce a correlation between dN and dS, regardless of whether the PS or MO approach is followed. When we consider the results of the simulation analysis, where we simulated data sets with no inherent dN and dS correlation and observed no correlation between the estimated dN and dS values, it seems the correlation between dN and dS observed in the empirical data is not introduced by the site definition method. There were no significant

reductions in the correlation between dS and ω between MO and PS models for any of the simulated data sets. In fact, in Simulation 6, we observed a significant increase in the correlation from the PS to the MO method, a result which counters our empirical findings. This could indicate that the inherent variation introduced during the dS and dN estimation process is large enough to influence the results. On the other hand, it could indicate that our empirical results where the MO method seemed to reduce the correlation between dS and ω when compared to PS method might not strictly be due to the method used to estimate sites, but perhaps due to some unknown biological factor.

The correlation between dS and ω is also greatly affected by the method used to estimate equilibrium codon frequencies. A significant reduction in the correlation was observed in most of the model pairs for all 6 organisms. The F61 method might implicitly incorporate properties inherent in the constitution of the coding sequence, such as extreme CUB. After reviewing the results of applying the $YN00$ method, it seems that CUB might be a plausible explanation as this phenomenon may be modeled more accurately by an F61 approach than by an F3x4 method, especially when the bias towards certain codons is very strong and F3x4 might not be precise enough. An interesting result was that of Simulation 2, in which the correlation between dS and ω was significantly reduced in all models when compared to the results from the data set of Simulation 1. The data in Simulation 1 were generated with all codon frequencies being equal, whilst the data set in Simulation 2 used unequal codon frequencies. This suggested that the positive correlation between dS and ω can be reduced by simply introducing more variation into the empirical data. Furthermore, the F61 method produced correlations between dS and ω which were again significantly reduced from the F3x4 models. This further hints that the F3x4 method might not pick up extreme CUB in alignments and that the F61 method is more accurate in these cases.

However, when we consider the worm and E. coli results for F61 models, it becomes apparent that the reduction in the correlation between ω and dS between the F3x4 and the F61 methods might, at least in part, be due to parameter estimate variance. The 95% confidence interval, assuming a Gaussian distribution, for the mean of the codon frequencies (which should be $1/61 = 0.01639344$) in

the global F61 human frequencies is: $(0.0002163433,\ 0.03257055)$, compared to the interval of the global F3x4 human frequencies: $(0.005450624,\ 0.02733627)$. Under the F61 approach, codon frequencies have high variance, and hence could introduce more variation into the dN and dS estimates. This in turn can reduce the correlation between these evolutionary rates, which might lead to a negative correlation between ω and dS.

The $M3$ model keeps dS constant in a gene while varying ω, or effectively dN among sites. Allowing for ω variation within a gene could lead to the reduced correlation between ω and dS as more variation is introduced into the ω parameter (which we estimate as $\omega* = p_0.\omega_0 + p_1.\omega_1 + (1 - p_0 - p_1).\omega_2$). When comparing the $M0$_F3x4_MO model to the $M3$_F3x4_MO model, all the organisms except human again showed reductions in the correlation between ω and dS. For all the organisms, the $M0$_F61_MO and $M3$_F61_MO models maintained the same direction of correlation between dS and ω as observed in the $M0$_F61_MO, but the magnitude of the correlation varied between the organisms. This could suggest that the introduced variation reduces the correlation (in other words the correlation is weaker) between dN and dS as well as between dS and ω. The results from Simulation 6 (Section 4.3.5) showed that, even though a strong positive correlation existed between dN and dS, we observed a strong negative correlation between dS and ω when fitting the $M0$ model variations.

We wanted to investigate whether the positive correlation observed between dS and ω in the empirical results was perhaps due to the $M0$ model not being able to effectively model ω variation between sites, since the $M3$ model significantly reduced the correlation between dS and ω for four of our six model organisms (yeast, fly, E. coli and worm) when compared to the $M0$ models. It seems that this might however not be the case since we observed a strong negative correlation between dS and ω for the $M0$ model in our simulation.

One finding that stood out in our simulation analysis was that, even though we generated data under strongly, positively correlated dN and dS values, we did not observe a positive correlation between dS and ω in any of these data sets (unless we explicitly introduced it such as in Simulation 3). This result was a counter to a finding by Drummond and Wilke [2008] in their simulation studies

(see Section 2.6.2).

We have shown how the authors had introduced a positive correlation between dS and ω by creating a non-linear dependence between dS and dN. Our simulations were run with a linear correlation between dN and dS, based on the fact that both rates depend on the mutation rate. After reviewing the simulation results, in which the method effects that we observed in the empirical data were not reproduced, we conclude that the assumption of a linear dependence between dN and dS might be too simple and is not representative of the actual relationship in the empirical data.

In closing, the effect of model parametrisation plays a role in the correlation patterns between evolutionary rates. It also seems that the correlation observed between dN and dS in the empirical data is not an artifact of methods used to estimate these values and could point to a biologically relevant factor. We now consolidate our findings with existing literature on the subject, incorporate our findings with those already presented on by Li et al. [2009] and Stoletzki and Eyre-Walker [2011], and place our findings' relevance with regards to the research objectives.

5.3 Conclusions and recommendations regarding research objective 1

We showed that one reason why models differ in their rate estimation and hence correlation patterns between evolutionary rates can be the way in which a site is defined. We showed that model pairs which are identical in their rate estimation method except for the way in which synonymous and nonsynonymous sites are determined can have opposite correlation patterns between dS and ω.

However, in our simulation results we found no significant reductions in the correlation between dS and ω from the PS method to the MO method. This could mean that, even though correlations between dS and dN were introduced to mimic the empirical data, the underlying biological factor which causes the

difference in the correlation between dS and ω from the PS methods to the MO methods in the empirical data could not be reproduced by our simulations.

These findings reiterate is that model choice is not an arbitrary decision and choosing an appropriate site definition depends on the purpose of the study. Bierne and Eyre-Walker [2003] recommends that when estimating rates of evolution such as dS and dN, the correct site definition depends on the application of the rate estimates. Under the assumption that synonymous mutations are neutral, dS can be interpreted as the average mutation rate across the three codon positions when using a MO approach.

They further argue that the MO definition could give misleading results when comparing rates between genes, but that the MO definition is more accurate when inferring positive selection, in other words when estimating ω for different sites in a gene. We conclude that model choice is still a difficult decision and that one needs to be aware of the resulting correlation patterns.

5.4 Conclusions and recommendations regarding research objective 2

It is known that ω does vary between sites in organisms and it has been found that allowing for ω variation when fitting models of evolution improves model fit [Yang et al., 2000, Bao et al., 2008, Huelsenbeck and Dyer, 2004]. We wanted to investigate whether the $M0$ model might be too crude to effectively account for when ω varied between sites, and that this was the cause of the positive correlation between ω and dS in the empirical results. When we fitted the $M3$ model to our empirical results, we observed a significantly reduced correlation between dS and ω from the $M0$_MO model to the $M3$_MO model in four of our six organisms (yeast, fly, E. coli and worm). Furthermore, the overall magnitude of the correlations was smaller in the $M3$ models. Allowing for site-to-site ω variation therefore reduces the correlation between dS and ω.

We also found in model fit analysis that the $M0$ model provided a better fit overall to the data sets, but, as discussed in Section 3.4, our pairwise alignments

may not provide enough information to effectively fit the $M3$ model to the data.

Our simulation results, however, showed that data which was generated with a strong dN and dS correlation and ω varied between sites did not show a positive correlation between dS and ω when variations of the constant rate $M0$ model were fitted to the data.

5.5 Conclusions and recommendations regarding research objective 3

Another factor that strongly influenced the correlation between dS and ω was the way in which codon frequencies were determined. We found that when a model pair was considered with one model employing the F3x4 approach and the other the F61 approach, the correlation was reduced from the F3x4 to the F61 approach in every instance. These results completed our third research objective, and showed that equilibrium codon frequency estimation methods affect the correlation between ω and dS.

The model fit analysis showed that an F61 parametrisation provided the best fit overall in all 6 species. We were interested to determine whether the results were due to the F61 approach's ability to implicitly model a biological trait. CUB plays a major role in evolutionary events [Rodrigue et al., 2008, Aris-Brosou and Bielawski, 2006, Duret, 2000] and would be more richly modeled by an F61 approach than an F3x4 approach, so we decided to investigate this phenomenon by fitting the Yang and Nielsen [2000] model (YN00) which explicitly incorporates CUB when calculating dN and dS.

The results of fitting the YN00 method were a strong negative correlation between dS and ω for all organisms. We therefore managed to show that when incorporating CUB into the dN and dS estimation process, a negative correlation between dS and ω was observed regardless of site definition and codon frequency estimation.

CUB has been found to be present in all six of our model organisms by various studies [Drummond and Wilke, 2008, Aris-Brosou and Bielawski, 2006, Yang

and Nielsen, 2008, Zhou et al., 2009, 2010], and seems to be aimed primarily at translational efficiency [Plotkin and Kudla, 2010].

Rodrigue et al. [2008], in an exploration of modeling CUB via codon frequency estimation methods, found that the F61 parametrisation was well suited but runs the risk of modeling other biological properties in a confounding manner, therefore CUB is better modeled by a simple codon frequency estimation method and a set of 60 specific CUB parameters (in other words a value for the preference of each codon in an alignment). Zhou et al. [2009] found that different genes have different optimal codons, which suggests the use of a F61 method per alignment, as opposed to a global F61 vector, but we warn that such an approach might be overly complex. In our model analysis, we found that the global F61 frequencies provided a better fit in five out of our six organisms (all except human).

We conclude that CUB is an important factor in evolutionary processes and should be modeled accordingly. The MO approach is especially sensitive to codon frequencies due to it incorporating these values in both the substitutions and sites calculations. We therefore recommend that when an F61 approach is followed, sufficient data must exist to accurately estimate the empirical codon frequencies.

5.6 Limitations of the study

The basic model fitted to our data was the $M0$ model [Yang et al., 2000], which is fundamentally the Goldman and Yang [1994] model. This is an old model parametrisation, but was used to evaluate the results of Drummond and Wilke [2008]. The Goldman and Yang [1994] and Muse and Gaut [1994] models were not developed to model selection acting on synonymous sites and it has been recommend by Yang and Nielsen [2008] that newer models should instead be implemented.

We were unable to emulate the empirical relationship between dN and dS in our simulation analysis. The simple linear relationship which we introduced between these rates did not lead to a positive correlation between dS and ω, which we

observed in most empirical data sets when applying the $M0$ model with a PS approach to estimating sites. We observed a negative correlation between dS and ω regardless of the methods used. Future work to establish the exact nature of the relationship between dN and dS could be insightful.

Our data consisted of pairwise alignments only and hence might not have been rich enough to accurately apply certain models to the data. The $M3$ model for instance showed a weak fit overall to the data sets, which is unexpected since it is known that site-to-site rate variation exists in empirical data. Using data that consists of multiple sequence alignments might be more suitable for rate estimation studies.

5.7 Conclusion

We have shown that the positive correlation observed between dS and ω by Wyckoff et al. [2005] and Drummond and Wilke [2008] is affected by model parametrisation. In particular, the way in which a site is defined and the method of calculating equilibrium codon frequencies not only reduced the positive correlation between dS and ω but in many cases this correlation became negative. We also found that explicitly incorporating CUB when estimating dN and dS often leads to a strong negative correlation between dS and ω. We found that the positive correlation between dS and ω was not due to a linear correlation between dN and dS in our simulation analysis. The empirical results showed that methods of estimating sites and methods of estimating codon frequencies had an impact on the correlation between dS and ω. We were unable to reproduce the result in our simulation analyses, but as discussed this was probably due to the simulations being unable to emulate the relationship between dN and dS. Allowing for ω variation between sites within a gene had an affect on the correlation between dS and ω in the empirical data, most notably an overall weaker correlation between these rates.

Our results extend previous work by Li et al. [2009] which showed that different models of evolution affect the positive correlation between dS and ω. We have shown which specific model parametrisations could be the cause of disparate

correlation patterns between models and have shown that the method used to estimate dN and dS could affect the correlation between dS and ω, regardless of the model fitted to the data.

In closing, we have shown that the positive correlation observed between dS and ω is affected by model parametrisation, specifically the codon frequency estimation method and site estimation method. However, our simulations show that the results do not apply to data in which a simple linear correlation between dN and dS exists and that further research into the cause of the correlation observed between ω and dS is needed. This leads to the conclusion that the positive correlation between dS and ω could possibly be artefactual in nature and should probably not be interpreted as biologically relevant until the relationship between dN and dS is better understood.

List of References

H. Akaike. *Information theory and an extension of the maximum likelihood principle.* Akademiai Kiado, 1973.

M Anisimova and C Kosiol. Investigating protein-coding sequence evolution with probabilistic codon substitution models. *Molecular Biology and Evolution*, 26 (2):255–271, October 2008.

S Aris-Brosou and J.P. Bielawski. Large scale analysis of synonymous substitution rates can be sensitive to assumptions about the process of mutation. *Genetics*, 378:58–64, 2006.

L Bao, H Gu, K.A. Dunn, and J.P. Bielawski. Likelihood-based clustering (LiBaC) for codon models, a method for grouping sites according to similarities in the underlying process of evolution. *Molecular Biology and Evolution*, pages 1995–2007, 2008.

N Bierne and A Eyre-Walker. The problem of counting sites in the estimation of the synonymous and nonsynonymous substitution rates: Implications for the correlation between the synonymous substitution rate and codon usage bias. *Genetics*, 165:1587–1597, 2003.

W Brown, EM Prager, A Wang, and AC Wilson. Mitochondrial DNA sequences of primates: tempo and mode of evolution. *Journal of Molecular Evolution*, 18:225–239, 1982.

F.S. Collins, M. Morgan, and A. Patrinos. The human genome project: lessons from large-scale biology. *Science*, 300(5617):286–290, 2003.

C. Darwin. On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. *New York: D. Appleton*, 1859.

C. Darwin and A.R. Wallace. *On the Tendency of Species to form Varieties; and on the Perpetuation of Varieties and Species by Natural Means of Selection*. Linnean Society of London, 1858.

A.M. Dean and GB Golding. The structural basis of molecular adaption. *Molecular Biology and Evolution*, 15:355–369, 1998.

W Delport, K Scheffler, and C Seoighe. Models of coding sequence evolution. *Briefings in Bioinformatics*, pages 1–13, October 2008.

W. Delport, K. Scheffler, G. Botha, M.B. Gravenor, S.V. Muse, and S.L. Kosakovsky Pond. CodonTest: modeling amino acid substitution preferences in coding sequences. *PLos Comp Biol*, 6:e1000885, 2010.

A Doron-Faigenboim and T Pupko. A combined empirical and mechanistic codon model. *Molecular Biology and Evolution*, 24:388–397, 2007.

D.A. Drummond and C.O. Wilke. Mistranslation-induced protein misfolding as a dominant constraint on coding sequence evolution. *Cell*, July 2008.

L. Duret. Evolution of synonymous codon usage in metazoans. *Current Opinion in Genetics & Development*, 12(6):640–649, 2002.

L Duret. tRNA gene number and codon usage in the c. elegans genome are coadapted for optimal translation of highly expressed genes. *Trends in Genetics*, 16(7):287–289, July 2000.

J Felsenstein. *Inferring Phylogenies*. Sunderland (MA): Sinauer Associates, 2004.

J Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–376, 1981.

R.A. Fisher. Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population. *Biometrika (Biometrika Trust) 10*, 4:507–521, 1915.

N Goldman and Z Yang. A codon based model of nucleotide substitution for protein coding DNA sequences. *Molecular Biology and Evolution*, 11(5):725–736, 1994.

M. Hasegawa, H. Kishino, and T. Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22 (2):160–174, 1985.

I. Hellmann, I. Ebersberger, S.E. Ptak, S. Pabo, and M. Przeworski. A neutral explanation for the correlation of diversity with recombination rates in humans. *The American Journal of Human Genetics*, 72(6):1527–1535, 2003.

J Huelsenbeck and K.A. Dyer. Bayesian estimation of positively selected sites. *Journal of Molecular Evolution*, 58:661–672, 2004.

Y. Ina. New methods for estimating the numbers of synonymous and nonsynonymous substitutions. *Journal of Molecular Evolution*, 40(2):190–226, 1995.

M. Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 6:111:120, 1980.

S.K. Kosakovsky Pond and S.V. Muse. Site-to-site variation of synonymous substitution rates. *Molecular biology and evolution*, 22(12):2375, 2005a.

S.L. Kosakovsky Pond and S.W. Frost. Not so different after all: A comparison of methods to detect amino acid sites under selection. *Molecular Biology and Evolution*, 22(5):1208–1222, February 2005.

S.L. Kosakovsky Pond and S. Muse. HyPhy: hypothesis testing using phylogenies. *Statistical Methods in Molecular Evolution*, pages 125–181, 2005b.

S.L. Kosakovsky Pond, S. Frost, and S.V. Muse. HyPhy: hypothesis testing using phylogenies. *Bioinformatics*, 21(5):676–679, 2005.

C Kosiol, I Holmes, and N Goldman. An empirical codon model for protein sequence evolution. *Molecular Biology and Evolution*, 24(7):1464–1479, March 2007.

J. Li, Z. Zhang, S. Vang, J. Yu, G.K.S. Wong, and J. Wang. Correlation between Ka/Ks and Ks is related to substitution model and evolutionary lineage. *Journal of Molecular Evolution*, 68(4):414–423, 2009.

B.Y. Liao and J. Zhang. Evolutionary conservation of expression profiles between human and mouse orthologous genes. *Molecular Biology and Evolution*, 23(3): 530, 2006.

H. Lindsay, V.B. Yap, H. Ying, and G.A. Huttley. Pitfalls of the most commonly used models of context dependent substitution. *Biology Direct*, 3(1):52, 2008. ISSN 1745-6150.

P Lio and N Goldman. Models of molecular evolution and phylogeny. *Genome Research*, 8:1233–1244, 1998.

T Miyata and T Yasunaga. Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its applications. *Journal of Molecular Evolution*, 16:23–36, 1980.

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*, 11:715–724, 1994.

M. Nei and T. Gojobori. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution*, 3(5):418, 1986. ISSN 0737-4038.

M Nei and S Kumar. *Molecular evolution and Phylogenetics*. Oxford University Press, 2000.

J.B. Plotkin and G. Kudla. Synonymous but not the same: the causes and consequences of codon bias. *Nature Reviews Genetics*, 12(1):32–42, 2010.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

N Rodrigue, N Lartillot, and H Philippe. Bayesian comparisons of codon substitution models. *Genetics*, 10.1534/genetics.108.092254, September 2008.

S. Schaffner and P. Sabeti. Evolutionary adaptation in the human lineage. *Nature Education*, 1(1), 2008.

N. Stoletzki and A. Eyre-Walker. The positive correlation between dN/dS and dS in mammals is due to runs of adjacent substitutions. *Molecular Biology and Evolution*, 28(4):1371, 2011.

N Sugiura. Further analysis of the data by Akaike's Information Criterion and finite corrections. *Commun Stat Theory Methods*, A7:13–26, 1978.

S Whelan, P de Bakker, and N Goldman. PANDIT: a database of protein and associated nucleotide domains with inferred trees. *Bioinformatics*, 19:1556–1563, 2003.

S Whelan, P de Bakker, E Quevillon, N Rodriquez, and N Goldman. PANDIT: an evolution-centric databse of protein and associated nucleotide domains with inferred trees. *Nucleic Acids Res*, 34:D327–D331, 2006.

G.J. Wyckoff, C.M. Malcom, E.J. Vallender, and B.T. Lahn. A highly unexpected strong correlation between fixation probability of nonsynonymous mutations and mutation rate. *Trends in Genetics*, 21(7):381–385, July 2005.

Z. Yang. Adaptive molecular evolution. *Handbook of Statistical Genetics*, 2004. doi: 10.1002/0470022620.bbc10.

Z. Yang. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24(8):1586, 2007. ISSN 0737-4038.

Z Yang. PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the BioSciences*, 13:555–556, 1997.

Z Yang and R Nielsen. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Molecular Biology and Evolution*, 17 (1):32–43, 2000.

Z Yang and R Nielsen. Mutation-selection model of codon substitution and their use to estimate selective strengths on codon usage. *Molecular Biology and Evolution*, 25(3):568–579, January 2008.

Z Yang, R Nielsen, N Goldman, and AM Pedersen. Codon substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, 155:431–449, May 2000.

T. Zhou, M. Weems, and C.O. Wilke. Translationally optimal codons associate with structurally sensitive sites in proteins. *Molecular Biology and Evolution*, 26(7):1571, 2009. ISSN 0737-4038.

T. Zhou, W. Gu, and C.O. Wilke. Detecting positive and purifying selection at synonymous sites in yeast and worm. *Molecular Biology and Evolution*, 27(8): 1912, 2010.

Appendix A

Result significance summary tables

Effect Investigated	Methods compared	Organism	Correlation in first method	Correlation in second method	Significant	p-value
Site definition	M0_F3x4_PS vs M0_F3x4_MO	human	0.1246	−0.08617	yes	10^{-4}
		mouse	0.04827	−0.04862	yes	10^{-4}
		fly	0.20115	0.0638	yes	10^{-4}
		yeast	0.11976	0.08628	no	0.0618
		worm	−0.02822	−0.05016	no	0.1635
		E. coli	0.10958	0.12038	no	0.3446
Site definition	M0_F61_PS vs M0_F61_MO	human	0.05887	−0.14669	yes	10^{-4}
		mouse	0.0467	−0.08039	yes	10^{-4}
		fly	0.18548	−0.00671	yes	10^{-4}
		yeast	0.07468	0.00517	yes	0.005
		worm	−0.62341	−0.62252	no	0.4721
		E. coli	−0.50812	−0.51639	no	0.3372
Cod. freq. estimation	M0_F3x4_PS vs M0_F61_PS	human	0.12282	0.05887	yes	0.0003
		mouse	0.04827	0.0467	no	0.4562
		fly	0.20115	0.18548	no	0.1685
		yeast	0.11976	0.07468	yes	0.0212
		worm	−0.02822	−0.50812	yes	10^{-4}
		E. coli	0.10958	−0.62341	yes	10^{-4}
Cod. freq. estimation	M0_F3x4_MO vs M0_F61_MO	human	−0.08617	−0.14669	yes	0.0005
		mouse	−0.08039	−0.10593	yes	0.0418
		fly	−0.00671	−0.06369	yes	0.0004
		yeast	0.08628	0.00517	yes	10^{-4}
		worm	−0.05016	−0.51639	yes	10^{-4}
		E. coli	0.12038	−0.62252	yes	10^{-4}

Table A.1: Result significance summary for site definition and codon frequency estimation methods (Sections 3.3.2 and 3.3.3).

Effect Investigated	Methods compared	Organism	Correlation in first method	Correlation in second method	Significant	p-value
CUB incorporation	$M0$_F3x4_PS vs YN00_CUB_PS	human	0.12282	−0.17922	yes	10^{-4}
		mouse	0.04827	−0.10593	yes	10^{-4}
		fly	0.20115	−0.06369	yes	10^{-4}
		yeast	0.11976	−0.11966	yes	10^{-4}
		worm	−0.02822	−0.26442	yes	10^{-4}
		E. coli	0.10958	−0.07547	yes	10^{-4}
site-to-site ω variation	M0_F3x4_MO vs M3_F3x4_MO	human	−0.08617	0.02688	yes	10^{-4}
		mouse	−0.04862	−0.05412	no	0.3557
		fly	0.0638	0.00574	yes	0.0003
		yeast	0.08628	0.01557	yes	0.004
		worm	−0.05016	−0.1302	yes	0.0002
		E. coli	0.12038	−0.0004	yes	10^{-4}
site-to-site ω variation	M0_F61_MO vs M3_F61_MO	human	−0.1466	−0.02297	yes	10^{-4}
		mouse	−0.08039	−0.06559	no	0.1587
		fly	−0.0067	−0.05088	yes	0.0044
		yeast	0.00517	0.00805	no	0.4443
		worm	−0.51639	−0.33904	yes	10^{-4}
		E. coli	−0.62252	−0.38035	yes	10^{-4}
Global codon frequencies	M0_F3x4-GF_PS vs M0_F3x4-GF_PS	human	0.1221	0.08277	yes	0.0162
		mouse	0.06127	0.03614	yes	0.0455
		fly	0.19788	0.1716	no	0.0537
		yeast	0.1008	0.06835	no	0.0594
		worm	-0.02239	-0.22828	yes	10^{-4}
		E. coli	0.15948	-0.28494	yes	10^{-4}

Table A.2: Result significance summary for CUB, site-to-site ω variation and global codon frequency methods (Sections 3.3.4 and 3.3.5)